Research Design in Clinical Psychology

FIFTH EDITION

ALAN E. KAZDIN

Pearson

Research Design in Clinical Psychology

FIFTH EDITION

Alan E. Kazdin
Yale University

PEARSON

Boston Columbus Indianapolis New York City San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor-in-Chief: Ashley Dodge
Program Team Lead: Amber Mackey
Managing Editor: Sutapa Mukherjee
Program Manager: Carly Czech
Development Editor: Christine Ambrose, iEnergizer Aptara®, Ltd.
Editorial Assistant: Casseia Lewis
Director, Content Strategy and Development: Brita Nordin
VP, Director of Marketing: Maggie Moylan
Director of Field Marketing: Jonathan Cottrell
Senior Marketing Coordinator: Susan Osterlitz
Director, Project Management Services: Lisa Iarkowski
Print Project Team Lead: Vamanan Namboodiri
Project Manager: Sudipto Roy
Operations Manager: Mary Fischer
Operations Specialist: Carol Melville
Associate Director of Design: Blair Brown
Interior Design: Kathryn Foot
Cover Art Director: Maria Lange
Cover Design: Lumina Datamatics, Inc.
Cover Art: Narcisse/Shutterstock
Digital Studio Team Lead: Peggy Bliss
Digital Studio Project Manager: Elissa Senra-Sargent
Full-Service Project Management and Composition: Garima Khosla, iEnergizer Aptara®, Ltd.
Printer/Binder: Courier Kendallville
Cover Printer: Phoenix

Acknowledgements of third party content appear on page 537, which constitutes an extension of this copyright page.

Library of Congress Cataloging-in-Publication Data

Names: Kazdin, Alan E., author.
Title: Research design in clinical psychology / Alan E. Kazdin, Yale University.
Description: Fifth edition. | Boston : Pearson, [2016] | Includes bibliographical references and index.
Identifiers: LCCN 2015048757| ISBN 9780205992089 | ISBN 0205992080
Subjects: LCSH: Clinical psychology—Research—Methodology.
Classification: LCC RC467.8 .K39 2013 | DDC 616.890072—dc23 LC record available at http://lccn.loc.gov/2015048757

2010048486

10 9 8 7 6 5 4 3 2 1

PEARSON

ISBN-10: 0-205-99208-0
ISBN-13: 978-0-205-99208-9

Dedicated to Nicole and Michelle

Brief Contents

Contents

Preface

The purpose of this text is to describe and explain research methods in clinical psychology but the issues and methods are relevant to other areas as well, such as counseling, educational, health, and school psychology, psychiatry, social work, and nursing. The topics within each of these areas span theory, research, and application. Consequently, many of the methodological challenges are shared. The text elaborates the methods of conducting research and the broad range of designs and practices for developing a sound knowledge base. The intended audiences are individuals who design and conduct research and who read research and wish to discern what can and cannot be concluded based on how that research was conducted.

Research in clinical psychology and other disciplines I have mentioned span well controlled laboratory settings as well as applications in clinic, community, and field settings where less control is allowed and the slings and arrows of everyday experience can interfere with drawing clear inferences. An in-depth understanding of methodology is of great importance because of the range of influences in clinical and applied research that can obscure the results. These influences cannot be used as an excuse for poorly designed research. On the contrary, the subject matter and the diverse ways in which research is conducted require a grasp of the underpinnings and nuances of design so that special arrangements, novel control conditions, and methods of statistical evaluation can be deployed to maximize clarity of our findings. Methodology, including the underlying tenets and specific practices, permit the combination of rigor and ingenuity as a defense against the multitude of influences that can obscure the relations among variables.

Clinical psychology encompasses a variety of topics including the study of personality, assessment and prediction of psychological functioning and positive adjustment, etiology, course, and outcome of various forms of psychopathology and their cognitive, social, and cultural neuroscience underpinnings, and the impact of interventions (treatment, prevention, education, and rehabilitation). Many issues of contemporary life have added to the range of research topics, as witnessed by the strong role that psychology plays in research on health, interpersonal violence, crime, trauma, homelessness, and substance use and abuse. Also, family life and demographic characteristics of the population have changed (e.g., increases in teenage mothers, single-parent families, blended families, and same-sex parenting; shift in population with more elderly who are physically active). Each of these and other changes has spawned rich areas of study directly related to understanding mental and physical health. Cultural and ethnic issues increasingly are recognized to play a central role in understanding variation in core psychological processes as well as adaptive and maladaptive functioning. These changes have made the substantive focus of psychological research in general very rich. Substantive foci and findings are very much intertwined to research methods and challenges to address these questions in an evolving society.

Methodology

Methodology as a broad overarching topic is divided in this text into five areas:

- Research Design,
- Assessment,
- Data Evaluation and Interpretation,
- Ethics and Scientific Integrity, and
- Communication of Research Findings.

These areas help organize many issues as they emerge in the planning and executing research from the developing the research idea, selecting methods, procedures, and assessment devices, analyzing and interpreting the data, and preparing the written report of the results. While there is an obvious sequence in planning and executing research, ethical issues in the treatment of participants and scientific integrity pervade all facets of methodology and before, during, and after a study is conducted. At each stage of research, underlying principles, options strategies, and guidelines are presented. Connections are made as well to convey how one facet of a study we have discussed (e.g., research design, assessment) influences another (e.g., ethical issues, communication of findings).

Many methods are covered as for example illustrated with major design options (e.g., true experiments, quasi-experiments, observational studies, single-case experiments for clinical use, qualitative research) and modalities of assessment (e.g., objective and projective measures, behavioral measures, neuroimaging). The goal is to convey the range of options so that one can move from hypotheses to design in different ways but also to consider strengths, weaknesses, and trade-offs in electing specific strategies.

Overall, methodology is addressed from multiple perspectives or levels of analysis. First, methodology is a way of thinking, problem solving, and approaching substantive questions. This focus emphasizes the commitment

to overarching principles that guide science and how we describe and explain data. Second and related, there are many specific concepts that direct our attention of what to consider and what facets of a study are likely to emerge as problems that interfere with obtaining clear information from our data collection. These concepts help us move from general abstractions of developing a research idea to considering the many conditions that form a study. Once these specific concepts are known, it is possible to evaluate virtually any scientific study. Also, the specific concepts we raise direct our attention to and anticipate a range of well-known biases and pitfalls.

Third, and as expected, methodology includes scores of specific practices from sampling, assigning subjects, matching, selecting data analyses, handling missing data, and so on. The text covers these in detail but in the process reflects back on underlying principles and specific concepts we are trying to address. It remains critical at each stage and with specific practices to keep in mind what we are trying to accomplish and why. That connection can open further options as to what we can do to strengthen the inferences we wish to draw from a study.

Finally, methodology is evolving within psychology and the sciences more generally. Of course, one can find stability in methodology. Random assignment of subjects to groups or conditions, when possible, is still wonderful. Yet, much of methodology continues to change. The standards for what constitutes a "good," "well controlled," and important study continue to evolve, the range of options for measurement, the use of technology and the Web in conducting studies and expanding beyond the usual range of participants, how participants in research subjects ought to be informed, treated, and protected, and what constitutes conflict of interest among investigators. The text covers many of the changes and the broader point that methodology is not at all static.

The text emphasizes the importance of methodological diversity in science and of course specifically psychological science. There are multiple methodologies in research and the focus, yield, and contributions of these vary. We usually learn in our training the importance of experiments based on groups, comparison of group differences, null hypothesis testing statistical evaluation, and so on. This is the emphasis of the present text because this is the dominant paradigm and students ought to master the strengths, methods, and weaknesses. There are other and methodologies and approaches; they are mentioned because they are important in their own right in relation to topics studied in clinical, counseling, educational, and other areas of psychology. Also, the methodologies convey and place into sharper focus many research practices we currently take for granted as the only paradigm for empirical science.

Methodological diversity is central to research for yet another reason. The methods we select among the many options available, how we frame the question, the groups we include, and the ways we decide to measure key constructs directly affect the answers we obtain. It is not the case that every answer to every question will change depending on our methods. Even so, it is important to understand that different answers can be readily achieved with different methodological tools and decisions. This is not a "problem." The different methods we use often reveal different facets of a phenomenon, a point illustrated as we present different methods.

Overview of the Text

Research includes several stages as an investigator moves from identifying the research question; translating that into a specific study; addressing potential sources of influence, which could obscure interpretation of the results, to obtaining, evaluating, and interpreting the data. Each of these and many intervening steps are points, and each decision has its own implications and trade-offs in terms of the final product. The principles of methodology tell us what we are trying to accomplish at the decision points and the procedures and practices help us concretely devise and implement the study.

The text describes and evaluates diverse research designs, methods of assessment, and many procedures and the rationale for their use. The goal is to be of concrete help to individuals who are designing studies and evaluating the studies that others have completed. This is not a recipe text with specific procedures and ingredients from which someone can simply select. Each practice serves a purpose, and it is important to understand what that is and what trade-offs there might be in selecting one practice versus another.

Chapter 1

This chapter provides an overview of the text and introduces the topic of research design as used in clinical psychology.

Chapters 2 & 3

Methodology includes arranging the circumstances of the study so as to minimize ambiguity in reaching conclusions. Many of the factors that can interfere with drawing clear conclusions from research can be readily identified. These factors are referred to as *threats to validity* and serve as the basis for why and how we conduct research—psychological research specifically but all scientific research more generally. Types of experimental validity and the factors that interfere with drawing conclusions serve as the basis for Chapters 2 and 3.

Chapter 4

The investigation begins with an idea that becomes translated into a specific question or statement. Yet, how does one develop an idea for research? Ideas come from many places. Chapter 4 discusses sources of ideas in different ways including the role of theory and types of research (e.g., basic, applied, and translational research). Also, the topics of what makes research interesting and important are discussed. Finally in this chapter is a guide for obtaining the research idea and then moving to the next steps to develop the study.

Chapter 5

The design or how conditions are arranged to test the hypothesis is an initial pivotal decision in moving from an idea to a study. Chapter 5 discusses different design options and arrangements including true-experiments and quasi-experiments and how they address the threats to validity. Also, group designs begin with deciding who will be the subjects or participants in research (e.g., college students, online sample from the Web, clinical population). This chapter considers different options and factors that guide participant selection and the critical role of diversity (e.g., ethnicity and culture) because of their influence on what is being studied.

Chapter 6

Control and comparison groups in a study obviously are pivotal and determine what can be concluded in a study. Different types of control groups, especially in the context of experiments and the evaluation of interventions, are presented. Each type of control or comparison condition is associated with the type of question the researcher wishes to ask but also may involve ethical and practical issues that guide the decision as well. Chapter 6 discusses several types of control and comparison groups and the considerations that dictate their use.

Chapter 7

A great deal of research is based on understanding variables that cannot be manipulated directly, as illustrated, for example, in the study of individuals with different characteristics (e.g., clinical disorders, experiences, and exposure to events—natural disasters such as hurricanes and human-made disasters such as war). Observational designs (case-control and cohort designs) in which individuals are selected and evaluated concurrently or longitudinally are presented in Chapter 7. These designs are quite powerful in identifying antecedents (e.g., risk factors to some outcome such as a mental or physical health problem, dropping out of school, criminality) and even possible causal relations. There are multiple design options, control procedures, and strategies to optimize the yield from designs in which variables of interest cannot be manipulated and controlled experimentally.

Chapter 8

Although experimental designs usually consist of group studies, causal inferences can be drawn from the study of individuals or a small number of individuals. Single-case experimental designs provide a methodology for drawing inferences that can be applied both to individuals and groups. The designs expand the range of circumstances in which can conduct evaluations, especially in circumstances where control groups are not available and one is interested in evaluating an intervention program. Chapter 8 presents special design and data-evaluation strategies that characterize single-case experimental research.

Chapter 9

The vast majority of research within psychology is within the quantitative tradition involving group designs, null hypothesis testing, assessment on standardized scales and inventories, and statistical evaluation in the form of null hypothesis testing. From a different tradition and approach, qualitative research methods alone but also in combination with quantitative research are enjoying increased use in psychology and social sciences more generally. Qualitative research is a scientifically rigorous approach and makes a special contribution to knowledge, usually by intensively studying a small number of subjects in depth. The goal is to capture the rich experience of individuals in special circumstances and to go well beyond the knowledge that can be obtained by questionnaires and fixed measures. Chapter 9 provides an overview of the qualitative research, conditions to which the designs are suited, and illustrations to convey the contribution to developing the knowledge base. Qualitative research, along with the prior chapter on single-case research, also places into perspective the dominant model of quantitative and hypothesis testing research and expands the range of options from those commonly used to address important research questions.

Chapter 10

The chapters now move from design strategies to measurement. Chapter 10 focuses on the underpinnings of assessment to establish key considerations in selecting measures for research and interpreting the measures that are presented in articles we read. Core topics of assessment are included such as various types of reliability and validity, the use of standardized versus nonstandardized measures, and assessment issues that can influence the conclusions one can reach from research. Useful strategies (e.g., selecting multiple measures, measures of different methods) and their rationale for improving research also are discussed.

Chapter 11

The varied options for measurement are discussed in Chapter 11. These options or assessment modalities include large families of measures such as objective, projective, observational, psychobiological measures, and other types as well. The chapter illustrates specific measures but is more concerned about conveying the different modalities and their strengths and limitations. In addition, the chapter encourages drawing from different types of measures in any one study to strengthen the conclusions that can be drawn.

Chapter 12

Special topics in assessment are covered in Chapter 12. The chapter begins by discussing ways on assessing or checking on the impact of experimental manipulations on the participant. These measures focus on whether the manipulation was perceived by or registered with the participants and are not primary outcomes or dependent variables. Assessment of the manipulation raises important issues to strengthen a study but also special considerations that can influence interpretation of the findings. Another topic in the chapter is measuring the practical or clinical significance of change that goes beyond the usual measures.

Chapters 13, 14, & 15

The next chapters turn to data evaluation. Null hypothesis and statistical testing serves as the dominant model in scientific research in social, natural, and biological sciences and of course including clinical psychology, counseling psychology, education, and other areas with basic and applied research questions. Mastery of the approach is essential. Chapter 13 evaluates the rationale of this approach and strategies to strengthen research within the tradition of null hypothesis testing. Common ways in which the results of research misinterpreted ("my results were almost significant; pretty please let me sort of say that they are significant") and failures to replicate the findings of others in light of statistical testing and binary decision making (significant or not) are also presented. Despite the dominance of null hypothesis testing, there is a long history continuing today firmly objecting to using the approach. Mastery of the approach requires knowing the objections and possible ways of addressing them. In addition, an alternative way of doing research (e.g., Bayesian analyses) is highlighted to convey another option from null hypothesis testing.

Data evaluation has many practical decision points related both to describe the sample and to draw inferences about the impact of the manipulation of interest. Chapter 14 discusses presentation of the data and using supplements to statistical significance testing (e.g., measures of strength of effect, confidence intervals) to elaborate the findings. Key decision points, multiple options, and sources of bias are highlighted in relation to such topics as handling missing data and deleting subjects from data analyses. Exploring one's data is also discussed to deepen one's understanding of findings but primarily as a guide to further hypotheses and studies. Chapter 15 focuses on interpretation of the findings of an investigation and common issues and pitfalls that emerge in moving from describing and analyzing the results to the interpreting of those results. This chapter also discusses so-called negative results, i.e., the absence of differences.

Chapters 16 & 17

Ethical issues and scientific integrity form the basis of Chapters 16 and 17, respectively. Although the topics overlap, I have treated them separately to permit their detailed treatment. For purposes of presentation, I have delineated ethical issues as the responsibilities of the investigator in relation to participants in research. The ethical issues chapter covers such key issues as deception, debriefing, invasion of privacy, informed consent and assent, withholding treatments, and presenting treatments of questionable effectiveness. Also, professional guidelines and codes along with federal regulations to guide protection of subjects are presented. Scientific integrity is delineated as the responsibilities of the investigator in relation to the research enterprise, science, and public trust. Issues that are covered include fraud, questionable practices that can distort findings, plagiarism, sharing of data, and conflict of interest, and jeopardizing the public trust. Here too there are professional guidelines and regulation to guide us. The chapters convey that ethical issues and scientific integrity are core features of research and emerge at the stage of developing the research proposal long before the first subject is run. In addition, ethics and scientific integrity are vibrant areas of activity in part because of greater public awareness of science and lapses in ethics or integrity but also because novel situations are emerging (e.g., "big data," findings that can be used for the public good or ill). These new situations raise the need for deliberation and new guidelines to ensure protection of subjects.

Chapter 18

Completion of a study often is followed by preparation of a written report to communicate one's results. Communication of the results is not an ancillary feature of research methodology. The thought and decision-making processes underlying the design of a study and the specific methods that were used have direct implications for the conclusions that can be drawn. Preparation of the report is the investigator's opportunity to convey the interrelation of the conceptual underpinnings of the study and how the methods permit inferences to be drawn about those underpinnings.

Chapter 18 discusses the written report and its preparation in relation to methodological issues presented in previous chapters. The special role that methodological issues and concerns play in the communication and publication of research is highlighted. Questions are provided to help guide the write-up of research on a section-by-section basis. Also, the journal review process and the different fates of manuscript will be of interest to those who develop research or read published articles.

Chapter 19

The text ends with closing comments that discuss the interplay of the five areas of methodology covered in prior chapters, namely, research design, assessment, data evaluation, ethical issues and scientific integrity, and communication of findings. The chapter conveys that substantive and conceptual issues and methodology are deeply intertwined. Methods used to study a phenomenon actually can contribute to the specific findings and conclusions. Consequently, the chapter underscores the importance of methodological diversity, i.e., the use of different methods (e.g., designs and measures) because different methods often elaborate different facets of a phenomenon of interest and produce different findings. The student who has completed and mastered the text will not need any simple, summary, nutshell rendition of how to develop and design the almost perfect study. Even so, at the very end of the chapter, there are simple guidelines for applying all that has been learned in a format that, hopefully, will assist any person designing his or her first study, or planning a project or grant.

New to the Edition

The revised edition of the text includes scores of additions and changes to reflect the evolving and dynamic nature of psychological science and methodology and ways of carrying out studies. Many such changes of this type addressed in this text, compared to prior editions, include greater attention to:

- How to develop a research idea and what makes a research study interesting and important;
- Use of technology and Web-based methods to conduct studies;
- Cultural and ethnic issues and how and why they are essential to integrate into research;
- Decision making in analyzing the results and points where bias often is introduced;
- Ethical issues and scientific integrity and their pervasive role in the research process from beginning to end;
- Publication bias, "negative" results, and current priorities related to replication; and
- Changes in the publication and communication of research that can affect both researchers and consumers of research.

I mentioned technology and its role in research design. Novel and emerging topics related to technology including secondary data analyses on a large scale, "big data," tracking individuals and connecting data (e.g., social network, GPS tracking of smart phones, monitoring purchases on the Internet), and the nature of publication of research (e.g., predatory journals, ghost authors) raise all sorts of new opportunities (e.g., assessment in real time, feedback to subjects in their everyday life) and problems. Several such topics have been expanded in the revised edition along with the many of the challenges (novel ethical issues, ways of reducing fraud).

Apart from additions, each chapter was revised and updated. An effort was made to retain classic references and references to leaders in statistics and methodology whose names ought to be known and recognized because of their roles in developing methods that we currently use. Also, many key topics of research were retained (e.g., moderators, mediators, and mechanisms) but updated in light of changes in research. Throughout the text examples are provided to illustrate key points. The examples draw from classic (old) but mostly new studies and from clinical and other areas of psychology.

For the illustrations of all components of methodology, I have drawn examples from natural, biological, and social sciences, in addition to psychological and clinical psychological research. The purpose in drawing from diverse fields is four-fold. First, psychology is recognized as a hub science, i.e., a field from which many other disciplines draw including education, medicine, law, economics, and public health. Our substantive findings as well as our methods routinely are drawn upon. This allows illustrations of what is important in methodology to connect with other areas of research. Many of the central issues and concerns specific to areas of this text (e.g., clinical, counseling, educational psychology) are common among many disciplines. Seeing a methodological issue or practice in different contexts can lead to better understanding as well as increase options for how we address the matter in our studies.

Second, disciplines often approach topics somewhat differently. For example, there are currently new and evolving guidelines regarding the use of placebos in medicine. The ethical issues and new guidelines developed to address them raise critical points in psychological research in relation to the various control and comparison groups we use (e.g., in evaluating the effects of psychotherapy or a community intervention to improve nutrition). In fact, guidelines and regulations often drawn for research in one area or discipline spill over into other areas as well. Seeing

emergent issues in other areas can deepen our understanding of many practices that are required in our research.

Third, psychologists (and scientists in general) increasingly are involved in collaborative arrangements with researchers from other disciplines. Indeed, many of the examples are drawn from just such instances. Thus methodologies from varied disciplines move back and forth to influence each other. Drawing examples from diverse disciplines helps to convey the methodological diversity, the range of options are available in research, and some of the advantages of collaborating to study phenomena of interest.

Finally, many fascinating examples from diverse areas can illustrate key points to bring methodology to life. For example, methodology is illustrated with examples on such topics as sports, sexual attraction, bullying in the schools, the effects of wine and religion on health, what stress can do to our immune system, cancer cures that could not be replicated, abstinence programs in the schools and their effects on sexual activity, racism and discrimination in research, interpersonal violence, and self-injury, so on. The purpose goes beyond the effort to make methodology engaging. Methodology is the core of key topics of our daily lives and is relevant. Stated another way, methodology is not merely a text on how to do or interpret studies. Methodology underlies the knowledge that we and others (e.g., policy makers, legislators) rely on to make decisions for ourselves, family members, or some group for which we have input or responsibility. Understanding the strengths and weaknesses of research and nuances are pivotal. Although there is an ivory tower feature of methodology, as scientists we are in the world and it is important to keep the relevance of what we do in mind as we design, complete, and write-up our research. Stated more dramatically but also accurately, methodology can be a matter of life and death and that point demands illustration and support. It is coming later in the text.

Although many examples draw on topics important to everyday lives that is not the only dimension on which current examples were selected. The range of research from laboratory to applied studies is addressed in separate ways. These include the role and importance of nonhuman animal studies and their contributions. Research projects designed to be a proof of concept, for example, convey how critical methodology is to see what can happen in principle. Also the range of translational research is discussed that include the extension of research from the laboratory to person or patient care ("bench-to-bedside" research) and from individual person care to community level intervention ("bedside-to-community" research).

This edition includes teaching aids for the reader and instructor. First, throughout the text, I have added tables to provide summaries and aids for the reader. When there are multiple points that require elaboration (e.g., how to increase power, types of relations among variables the investigator may wish to study), it is easy to lose sight of the key points. The tables are useful study guides once the individual entries have been elaborated. Second, at the end of each chapter there is a chapter summary to assist the reader in reviewing key concepts. Third, there is a list of readings included at the end of the text that directs the interested reader to more in-depth presentations of topics; this listing is organized by chapter. Finally, a Glossary is included at the end of the text to centralize and define briefly terms introduced throughout the chapters. Special terms italicized within the text are usually covered in the glossary as well. Although the text is not overabundant in terminology, there is value to providing a quick reference to terms and practices.

REVEL™

Educational technology designed for the way today's students read, think, and learn

When students are engaged deeply, they learn more effectively and perform better in their courses. This simple fact inspired the creation of REVEL: an immersive learning experience designed for the way today's students read, think, and learn. Built in collaboration with educators and students nationwide, REVEL is the newest, fully digital way to deliver respected Pearson content.

REVEL enlivens course content with media interactives and assessments — integrated directly within the authors' narrative — that provide opportunities for students to read about and practice course material in tandem. This immersive educational technology boosts student engagement, which leads to better understanding of concepts and improved performance throughout the course.

Learn more about REVEL http://www.pearsonhighered.com/revel

Available Instructor Resources

The following resources are available for instructors. These can be downloaded at http://www.pearsonhighered.com/irc. Login required.

- **PowerPoint**—provides a core template of the content covered throughout the text. Can easily be expanded for customization with your course.
- **Instructor's Manual**—includes a description, in-class discussion questions, a research assignment for each chapter.
- **Test Bank**—includes additional questions beyond the REVEL in multiple choice and open-ended, short and essay response, formats.
- **MyTest**—an electronic format of the Test Bank to customize in-class tests or quizzes. Visit: http://www.pearsonhighered.com/mytest.

Acknowledgments

Several persons have contributed to the thrust and focus of this text over the last several years. It is usually gracious for an author to convey to the reader that any errors that remain in the text after extensive input from others are his or her responsibility alone. That is not how I feel. For errors, short-sightedness, limitations, and non-brilliant ideas in this text, I hold most people in my life responsible! My early upbringing in the forest, in utero fast foods fed to me over which I had no control, a maladaptive polymorphism here and there, and crushing judgmental frowns by an influential high school teacher or two are just some of the influences that account for the lapses that the reader may find in my thinking. Also, my peer group in the other incubators in the maternity ward the few days after my birth were not exactly positive influences—many other infants were slackers (they slept most of the time); others seemed to whine (e.g., cry when they did not get fed or changed). In that environment, I did the best I could but the limitations cannot be eliminated. Who knows what of those influences entered this text.

As to the positive influences, I have been blessed with remarkable colleagues and students who through direct discussion or exemplary work have inspired me to think about methods, how important they are, and what they can accomplish at their best. Insofar as this revision excels and is helpful, interesting, or important, I am pleased to share the credit. A few mentors deserve especial credit for their influence and include Richard Bootzin, Donald Campbell, and Lee Sechrest. Long ago but also in an enduring way, they inspired my interest in methodology and its importance. Fast forward to now, graduate and undergraduate students at Yale University who have taken course on the topic of this text also have provided detailed input and comment. I am especially grateful to those few students who did not demand refunds for the text halfway into the course.

Finally, although many years have passed since my dissertation, I owe a special debt of gratitude to my dissertation committee. In addition to the laugh track they played after I summarized my study at my dissertation oral exam, committee members made subtle, nuanced comments that linger in their influence on me (e.g., "Alan, find another career." "Research isn't for everyone." "When we said, 'use a pretest,' we did not mean omit the posttest.") These pithy comments raised the prospect that understanding methodology may be rather important. (Not wanting to be identified with my study, all my committee members entered the Dissertation Committee Witness Protection Program immediately after my oral exam, and unfortunately cannot be identified by their original names. But, thank you "Cody," "Billie Sue," "Thaddeus," and most of all the chair of my committee, "Mygrane." I am grateful to you all wherever you are.)

Several sources of research support were provided during the period in which this text was written. I am pleased to acknowledge grants from the National Institute of Mental Health, The Humane Society of America, The Laura J. Niles Foundation, Yale University, and a generous donor who wishes to remain anonymous. Needless to say, the views expressed in this text do not reflect the views of any agency that has provided research support nor, for that matter, the agencies that have not provided support.

Alan E. Kazdin

This page intentionally left blank

About the Author

Alan E. Kazdin, PhD, is Sterling Professor of Psychology and Professor of Child Psychiatry at Yale University. Prior to coming to Yale, he was on the faculty of the Pennsylvania State University and the University of Pittsburgh School of Medicine. At Yale, he has been Chairman of the Psychology Department, Director of the Yale Child Study Center at the School of Medicine, and Director of Child Psychiatric Services at Yale-New Haven Hospital.

Kazdin's research has focused primarily on the treatment of aggressive and antisocial behavior in children (inpatient and outpatient) and parent, child, and contextual influences that contribute to child dysfunction and processes and outcome of child therapy. His work has been supported by the National Institute of Mental Health, the William T. Grant Foundation, the Robert Wood Johnson Foundation, Rivendell Foundation of America, Leon Lowenstein Foundation, the Humane Society of America, the Laura Niles Foundation, and Yale University. His work on parenting and childrearing has been featured on NPR, PBS, BBC, and CNN, and he has appeared on Good Morning America, ABC News, 20/20, and Dr. Phil.

Kazdin has been editor of various professional journals (*Journal of Consulting and Clinical Psychology*, *Psychological Assessment*, *Behavior Therapy*, *Clinical Psychology: Science and Practice Current Directions in Psychological Science*, and *Clinical Psychological Science*). He has received a number of professional awards, including the Outstanding Research Contribution by an Individual Award and Lifetime Achievement Awards (Association of Behavioral and Cognitive Therapies), Outstanding Lifetime Contributions to Psychology Award and Distinguished Scientific Award for the Applications of Psychology (American Psychological Association), and the James McKeen Cattell Award (Association for Psychological Science). In 2008, he was president of the American Psychological Association.

Kazdin's 700+ publications include 49 books that focus on methodology, interventions for children and adolescents, parenting and child rearing, cognitive-behavioral treatment, and interpersonal violence. Some of his recent books include:

Single-Case Research Designs: Methods for Clinical and Applied Settings (2nd ed.)

Methodological Issues and Strategies in Clinical Research (4th ed.)

Parent Management Training: Treatment for Oppositional, Aggressive, and Antisocial Behavior in Children and Adolescents

The Kazdin Method for Parenting the Defiant Child: With No Pills, No Therapy, No Contest of Wills (with Carlo Rotella)

Behavior Modification in Applied Settings (7th ed.)

Evidence-Based Psychotherapies for Children and Adolescents (2nd ed.) (with John R. Weisz)

Violence Against Women and Children: Volume I: Mapping the Terrain. Volume II Navigating Solutions (with Jacqueline W. White and Marry P. Koss)

This page intentionally left blank

Chapter 1
Introduction

Learning Objectives

1.1 Justify the indispensability of science

1.2 Report some of the roadblocks in our study of science

1.3 Examine the methodologies that govern scientific research

1.4 Analyze some of the key concepts that guide scientific thinking and problem solving

1.5 Discuss the importance of Semmelweis's usage of a scientific way of thinking to solve a problem.

Science is the study of phenomena through systematic observation and evaluation. A body of knowledge in a given area is accumulated through agreed-upon methods about how to obtain and verify that knowledge. Science also is a special way of knowing. It relies on information from our experience and encounters with the world. Yet, it is a more formal way of understanding and evaluating that experience.

Key processes and characteristics of science are the use of:

- Generating theory or conceptual explanations of the phenomena of interest
- Proposing hypotheses to test these explanations
- Collecting data under conditions and special arrangements (e.g., experiments, natural situations)
- Evaluating the data to draw inferences about the hypotheses

The processes or steps do not need to flow in that order at all. We might systematically observe a relation that we did not expect. For example, women who immigrate to a country and have their children are more likely to have a child with autism than are women who are from the country (i.e., are already there) (Lehti et al., 2013). That finding has been replicated; so for the moment, let us assume this is reliable. That finding itself seems odd and not easy to explain. We now try to understand this.

- What about these mothers or families could explain the finding?
- Are less healthy moms the ones who migrate?
- Are they just as healthy but the stressors associated with migration (e.g., perhaps fleeing war zones) lead to many birth complications?
- Does migration temporarily lead to deficiencies in diet that somehow are involved?
- Are there new pathogens (bacteria, viruses) in the new country to which their immune systems have not accommodated?
- Where to begin?

The answer is developing a plausible explanation (theory) and now testing it. Age and income of the parents or complications in delivery of the child did not explain the effect. We turn to other possible explanations and also see if there is related research that could help. We know that low intake of folate (B9: a water-soluble B vitamin found in leafy green vegetables) increases risk of autism and that giving moms folate supplements decreases incidence of autism. Yet, diet is only one possibility, and we do not know from the immigrant study whether there were any dietary differences. We have our research tasks cut out for us but how wonderful it will be once we understand because then we can be the most helpful to prospective parents to reduce or eliminate the higher risk of autism. In that process, we are likely to learn about other disorders and the broader impact of parent practices before and during pregnancy and later child development. Perhaps armed with a fuller explanation, we can greatly reduce the rates of autism among mothers at risk. But this all began with an observed relation and that enters us into the key processes that characterize scientific research.

1.1: Why Do We Need Science at All?

1.1 Justify the indispensability of science

This is a good question. Four reasons can make the case for why we need science.

1.1.1: Rationale

Here are the four reasons that make the case for why we need science.

***First**, we need consistent methods for acquiring knowledge.*

There are many sciences, and it would be valuable, if not essential, to have the principles and practices consistent. We would not want the criteria for what "counts" as knowledge to vary as a function of quite different ways of going about obtaining that knowledge. This consistency is more important than ever because much of research on a given topic involves the collaboration of scientists from many different fields to address a question. They must speak the same language, share the same underlying values about how to obtain knowledge, and agree on procedures and practices (e.g., statistical evaluation, reporting data that do and do not support a particular hypothesis). Consistency also is critical within any given scientific discipline. For a given science (e.g., psychology), we would want consistency throughout the world in what the standards are for obtaining scientific knowledge—the accumulation of knowledge from all individuals in a given field requires this level of consistency. Science "says" essentially these are our goals (e.g., describe, understand, explain, intervene where needed, possible, and desirable) and these are our means (use of theory, methodology, guiding concepts, replication of results). Science is hardly a "game" because so many of the tasks we have are serious. Yet there are rules, and there are enormous benefits from following them among all sciences and scientists.

***Second**, science is needed to identify, detect, isolate, and reveal many of the extremely complex relations that exist in the world.*

Casual observation cannot identify the complexities that we study in science. Science uses special controlled arrangements to isolate influences that are otherwise difficult, if not impossible, to detect in everyday life. Also, science often relies on special methods of assessment that extend well beyond what our senses could reveal from normal observation. The complexities of our findings that require this special scrutiny that science provides are easily conveyed by examples from the natural and social sciences.

Consider questions and answers that scientific methods were needed to address:

- What is near the boundary of our universe? Well for starters, a galaxy (system of millions or more stars held by gravitational attraction) has been identified that is over 13 billion light years away.
- How did dinosaurs become extinct? Approximately 66 million years ago (give or take 300,000 years), a huge asteroid (15 kilometers or over 16,400 yards wide) crashed into the earth (near Yucatan, Mexico) and led to the extinction of more than half of all species on the planet, including the dinosaurs. The material blasted into the atmosphere would have led to a chain of events leading to a "global winter."
- Are male and female interactions and behaviors influenced by a woman's menstrual cycle? The place a woman is in her menstrual cycle apparently has effects on her behavior (e.g., selection of clothing, gait when walking, and the type of male that seems attractive, and how men respond to all of this). All of this is out of consciousness but conveys a dynamically changing interaction influenced in part by ovulation cycles.
- Exercise can greatly improve mental health, but how? Consider depression as one example. Exercise increases a protein in the brain (hippocampus) that helps the development of neuron and synapses (neurogenesis) and in the process reduces symptoms of clinical depression. These are the changes also made when antidepressant medication is used as the treatment.
- Do early harsh environments for children (e.g., exposure to violence, enduring stress, corporal punishment) have any long-term effects? Yes, they can have many including enduring impairment on the immune system (ability to ward off infection and inflammation), and that is considered to be the reason that such children have premature deaths from serious disease much later in adulthood.

This random-like sample of findings (each from a larger literature of multiple studies) is hardly the tip of the iceberg, and many findings you already know from your studies fit into the category, namely, they would be difficult or impossible to discern from casual observation. The complex findings required very special observation procedures under special arrangements and often using special math or statistics. The conclusions I list are not discernible by everyday observation. If you said, you knew all along there was a galaxy at the boundaries of our universe, what's the big deal? Or that of course exercise changes a specific protein in that area of the brain, you are among a very small group.

***Third**, whether the relations are complex or not, for many questions of interest, we need extensive information (a lot of data) to draw conclusions.*

How to obtain that information (assessment, sampling) requires very special procedures to yield trustworthy results. For example, how many individuals in community samples (i.e., in everyday life) experience some form of psychiatric disorder? To answer this, we need a large sample, a representative sample, and special procedures (e.g., use of measures known to be consistent with the information they provide and to reflect the phenomenon of interest). Approximately 25% of the population in the United States at any given point in time meet criteria for one or more psychiatric disorders (Kessler et al., 2009; Kessler & Wang, 2008). That kind of information cannot be obtained from casual observation or individual experience. (In fact, based on my informal assessment from a recent family reunion, I had the rate closer to 80%.) We need large data sets and systematically collected data to address questions, and science is needed to provide the information and in a trustworthy, transparent, and replicable way.

***Finally**, we need science to help surmount the limitations of our usual ways of perceiving the environment and extracting conclusion.*

There are many sources of subjectivity and bias along with limitations in our perceptions that interfere with obtaining more objective knowledge, i.e., information that is as free as possible from subjectivity and bias. How we perceive and think is wonderfully adaptive for handling everyday life and the enormous challenges presented to us (e.g., staying out of danger, finding mates and partners, rearing children, adapting to harsh and changing environments, meeting the biological needs of ourselves and family—it is endless). Our evolution spanning millions of years has sculpted, carved, sanded, and refined these skills, so I am not dismissing them here. Yet, those very adaptive features actually can interfere, limit, and distort information presented to us and do so by omission (our perception omits many facets of experience that we do not detect well) and by commission (we actively distort information on a routine basis).

1.2: Illustrations of Our Limitations in Accruing Knowledge

1.2 Report some of the roadblocks in our study of science

The goal of science is to build a reliable (consistent, replicable) body of knowledge about the natural world (physical, biological, psychological). Some limitations emerge that are merely part of being human that we need to address and surmount. Here is a brief sample, beginning with some you already know well.

1.2.1: Senses and Their Limits

Limitations of our senses including vision, hearing, and smell are familiar examples to convey how we are very selective in the facets of reality that we can detect. We consider what we see, hear, and smell to represent reality, i.e., how things are. In a way what we see, hear, and smell *are* reality. Yet, they are very selective. We do not see very much of the electromagnetic spectrum. We see what is called (and is amusingly self-centered) "the visible spectrum." Actually, it is not *the* visible spectrum but is *a* visible spectrum, because it is defined as that part of the spectrum that the human eye can see. We see wonderful things all of the time, people, colors, sky, sunset, and methodology texts, all the while knowing intellectually at least that we do not see it all. We do not see many parts of the spectrum (e.g., infrared, ultraviolet). Other animals (e.g., birds and bees and many other insects) see part of the spectrum we do not see that helps with their adaptation (e.g., identifying sex-dependent markings of potential mates that only are visible in ultraviolet light). The same holds true for sounds and smells; many nonhuman animals have senses that evaluate different parts of the world from those we can experience. Many animals can hear sounds that we do not hear (e.g., dogs, elephants, pigeons) and have a sensitivity to smell that vastly exceeds our own sense of smell (e.g., bears, sharks, moths, bees). More generally, many nonhuman animals trump our vision, hearing, and smell or have differences that are not better (more sensitive) or worse but just different (e.g., seeing different parts of the electromagnetic spectrum).

These examples are intended to make one point: as humans we see one part of the world and that is quite selective. The picture we have of what "is" omits piles of things that are. (As I write this paragraph, I am listening to a lovely tune on a dog whistle—I cannot really hear it of course, but the piece is written by Fido Johnson who has been called the Mozart of dog composers.) So one reason for science is to overcome some of the physical limitations of our normal processing of information. Much of what we want to know about and see cannot be seen by our ordinary capacities (our senses).

1.2.2: Cognitive Heuristics

Leaving aside physical limitations on seeing, smelling, and hearing the world, more persuasive arguments of the need for science come from many areas of cognitive psychology. These are more persuasive in the sense that when we look at experience well within our sight and capacities

of our senses we still may have enormous limitations in how we process that information. You already know the everyday expression, "seeing is believing;" psychological research has provided considerable support for the additional claim, "believing is seeing." We process the world in special ways and various cognitive processes have been well studied. These processes can and often do systematically distort and lead us to make claims and inferences that do not reflect reality, as revealed by less or unbiased means.

There are several characteristics of normal human functioning that reflect how we organize and process information. They are referred to as *cognitive heuristics* and are processes out of our awareness that serve as mental shortcuts or guides to help us negotiate many aspects of everyday experience (Kahneman, 2011; Pohl, 2012). The guides help us categorize, make decisions, and solve problems. The heuristics emerge as "bias" when we attempt to draw accurate relations based only on our own thoughts, impressions, and experience. There are several heuristics (as covered in the cited references).

Consider the *confirmatory bias* as an example of one cognitive heuristic. This heuristic reflects the role of our preconceptions or beliefs and how those influence the facets of reality we see, grasp, and identify. Specifically, we select, seek out, and remember "evidence" in the world that is consistent with and supports our view. That is, we do not consider and weigh all experience or the extent to which some things are or are not true based on the realities we encounter. Rather we unwittingly pluck out features of reality that support (confirm) our view. This is particularly pernicious in stereotypes, as one case in point. Thus, if one believes that one ethnic group behaves in this or that way, or that people from one country or region have a particular characteristic, we will see the evidence that is supportive—the supportive evidence is more salient in our mind and memory. Counter-evidence does not register as salient or if and when it does is dismissed as an exception.

1.2.3: Additional Information Regarding Cognitive Heuristics

Consider one of many lamentable stereotypes that has been part of our culture, namely that obese people are jolly, not based on research at all and even refutable. Furthermore, consider the following: you see eight pensive, mildly mournful obese individuals during your day and two other outgoing, smiling, and jolly obese individuals that same day. Our conclusion would not be (from casual observation) that a few obese people are jolly, or roughly 20% are. If one believes obese people tend to be jolly, the confirmatory biases would draw on the two as, "Aha, I knew it, no surprise here the group is jolly, but of course there are exceptions" or "those nonjolly ones probably just were having a bad day." You might even blurt out a cliché to even provide further confirmation by noting, "the exception proves the rule." The technical term for all of this processing is "normal," and other terms might apply too (e.g., stereotyping, prejudice, discrimination). Yet the coding of information is out of awareness completely but clearly guides our interpretation of reality. We need science in part to surmount such influences.

Of course it is quite a legitimate empirical (scientific) question to ask, for example, whether obese people are jolly, jollier than nonobese people, handle situations (e.g., pain, stress) with more positive outlooks, and so on. No single study could answer these, but it is interesting to note in passing that a gene associated with obesity also is related to depression. Obese individuals tend to have slightly lower rates of depression in light of a genetic influence that apparently influences both obesity and depression (Samaan et al., 2013). This finding is not the same as showing that obese individuals are walking around laughing and engage in inappropriately cheery behavior (e.g., at funerals). And we do not know what level of obesity (how much overweight, at what age, for how long) provides the limits of this finding. The point is that we cannot trust our perceptions in light of a confirmatory bias. And this is merely one form of cognitive bias in which our view, perceptions, and conclusions systematically depart from what the data in the world would show if the bias could be controlled in some way. There are many others that lead us to overestimate one possibility (e.g., being struck by lightning) or to underestimate others (e.g., being in a car accident while texting or talking on a phone while driving).

Cognitive heuristics are not the only set of influences that guide our perception. Our motivation and mood states can directly influence how and what we perceive of reality (Dunning & Balcetis, 2013). Both biological states (e.g., hunger, thirst) and psychological states (e.g., mood) can directly guide how reality is perceived. This is sometimes referred to as *motivated perception* or *wishful perceiving*. For example, when one feels threatened or angry, one is likely to see others as holding a weapon rather than a neutral object (Baumann & DeSteno, 2010). That is, the "reality" we perceive is influenced by us as a filter, and we are changing in biological and psychological states that have impact on what we see, hear, and recall.

1.2.4: Memory

Other examples illustrate how our normal processing of information influences and distorts. Consider a few facets of memory, a key topic within psychology. Memory refers to the ability to recall information and events,

although there are different kinds of memory and ways of studying them. As humans we believe (and are often confident) that our memory *records* reality but research very clearly shows that we *recode* reality (Roediger & McDermott, 2000). That is, more often than not we do not recall things as they have happened. And this has come up in many contexts.

***First**, as we consider stories of our past (e.g., childhood, high school years) little details and sometimes larger ones get filled in and become part of our remembered story.*

Our memory draws on information for experience of the external world, but these are filled in with internal processes (e.g., imagination, thought). As we recount the story, we cannot make the distinction between what in the story actually happened and what did not. *Reality monitoring* is the name for a memory function that differentiates memories that are based on external (the world) versus internal (one's own thoughts, perceptions) (Johnson, 2006). Thus, I can separate my imagined phone call from the Nobel committee (last night's dream) from reality (the phone call I actually received yesterday from my dry cleaner—pick up my shirts or they will be thrown out). Errors occur when that distinction is not made, and that is a function of several things including how vivid the imagined events are and how consistent they are with the external stimuli. We develop a story or scheme of an event or what happened and fill in details where and as needed, and when we recall the event cannot always distinguish the source. I have a vivid memory of something at home when I was 6 months or so old. This is a picture of where I was sitting, who entered the room, and so on. More likely, I was told related stories about this event many times and now subjectively I am certain I can recall this. I can recall this—but it is as likely as not, the event was registered on my memory by the stories and not by my direct recall of the event as it occurred, if it occurred at all.

***Second** and related, the notion of false memories has been in public as well as scientific literature.*

The interest emerged from the experiences of many clients in therapy who, over the course of treatment, newly recalled childhood experience of abuse that was brought out during the course of therapy. In fact, in several cases it looks as if the memories were actually induced by the very process of therapy. This does not mean of course that all, most, or any given recollection of abuse is false, but we know that some are and that is just enough. Research has moved to study false memories—can we induce them in stories, memory tasks, and laboratory studies (e.g., Brainerd & Reyna, 2005)? Yes, in experiments we can even implant them. And when people recall material in the experiment, often false memories (things that did not occur at all) in fact are recalled and mixed with those that have occurred.

***Finally**, consider recall used heavily by the courts in legal proceedings.*

In jury trials, the most persuasive type of evidence is eye-witness testimony. Juries are persuaded by a witness on the stand saying he or she saw the defendant do this or that and perhaps even identified the defendant out of a line-up as the perpetrator. The reliance of eye-witness testimony makes forensic psychologists want to jump out of their basement windows because there is now rather extensive research showing that this type of testimony is the most unreliable form of evidence and is responsible for sending more innocent victims to prison than any other form of evidence (Wells & Loftus, 2013). Well beyond our discussion are multiple findings that show that who is identified as the alleged criminal depends on how questions are presented to a witness, how the line-up of possible suspects is presented (one at a time, all together), the time between witnessing the event and recall, and so much more. Now rather extensive research not only has shown that eye-witness testimony is fairly unreliable, but also the many variables that influence what people recall and its accuracy. In short, coding and recalling experience, even when vivid and something in which we are very confident, may not represent what has happened. We need more reliable tools to codify current and past experience that surmounts some of our normal recall and other limitations.

1.2.5: General Comments

Several facets of perception, thoughts, and emotions influence how we characterize the world, although I mentioned only a small sample (e.g., only one cognitive heuristic although there are several; only a few areas of memory research including reality monitoring, false memories, and eye-witness testimony while omitting others). The point was just to convey that as humans we have limitations that can readily influence conclusions we reach. These limitations can have little impact (e.g., details regarding who was at a social event last month and who drank and ate what) or enormous impact (e.g., who goes to jail or receives the death penalty). Also, we negotiate life rather well, do not bump into buildings or each other when walking down the street, put on our clothing correctly most days, and say "hi" rather than "goodbye" when we first encounter a friend or colleague during the day. So we should not distrust our senses, cognition, and affect. Accumulating scientific knowledge is another story.

For developing a knowledge base of how the natural world is, the limitations I have illustrated convey how

essential it is to develop means to counter normal experience, perception, memory, and the like.

- The challenge is as follows: we know we have limitations in our perception and hence in our ability to acquire unbiased knowledge without some systematic set of aids.
- The paradox: we ourselves, with these imperfections, have the responsibility of developing those aids (methods) to surmount those limitations.

> *Methodology* is the broad label for principles, practices, and procedures we have devised to help overcome or minimize biases that can obscure our knowledge of what the world is like.

Methodology is invented by people and is hardly perfect or flawless. As a human endeavor, most human characteristics and imperfections (e.g., greed, fraud, distortion) are or can be involved along with so many of our ideal characteristics (e.g., search for true knowledge, cooperation, interest in helping others, understanding our place in the universe).

Think of science as a way of knowing filled with checks and balances. For example one check, arguably the most important, is repetition of findings by other investigators. This repetition of findings is referred to as *replication*. For example, if I find an amazing result and no other investigator can reproduce (replicate) that after many excellent tries, my finding is suspect. I am not necessarily suspected of anything odd, but the finding is not reliable. Perhaps the finding depended on something none of us knows about or occurred by chance, fluke, or a bias I did not detect or control. At this moment in our discussion, the reason does not matter. But we have to say that my finding is not to be taken as a reliable finding and we go on. Perhaps some people replicate my finding but others do not. This suggests there may be some other condition or circumstance (e.g., perhaps some characteristic of the participants? Perhaps how the experimental manipulation is conducted?) that influences whether the finding is obtained. More work is needed to reveal if that is true. Yes, if my study cannot be replicated, that is annoying at the moment, but we are committed to the process and the last thing any scientist wants is to squeeze in "false knowledge," i.e., findings that do not hold up across investigators, laboratories, and time.

We will say more about replication and all the things failure to replicate can mean but for now, methodology is the answer developed by humans to provide the best information we can, so that it can be believed, accumulated, relied on, and repeated.

- Methodology does not eliminate bias and problems, and so a great dose of humility about the process is just wise.
- Methodology is dynamic and constantly developing as we learn novel ways in which bias may enter, novel ways to control that, and better measures of everything we do to monitor how a study is conducted and to measure constructs we care about with greater precision.
- Methodology is evolving, improving, and correcting sources of bias or influences that can interfere with obtaining knowledge.
- Methodology can contribute enormously to our lives leaving aside the lofty goals of developing our knowledge base.

I believe you personally value, if not love, methodology or will someday, even though you may not know it yet. (Methodology is love at last sight rather than first sight.) One hopes that now or in the future you or one of your relatives will not require treatment (medical, psychological) for a seriously debilitating condition (e.g., cancer, stroke, major depression, posttraumatic stress disorder). Yet for these and many other conditions, there are evidence-based interventions that can really help. Those interventions were developed and evaluated with sound research methods using all sorts of principles, practices, and procedures we will discuss in this text. Rarely does casual observation provide the means of identifying effective interventions. Methodology allows us to obtain the needed knowledge and that knowledge often saves lives and makes lives better—our own personal lives and those whom we love and like. Do you like methodology now? Me too.

1.3: Methodology

1.3 Examine the methodologies that govern scientific research

The topic of this text is methodology of psychological science with particular emphasis on clinical psychology, counseling, education, and social sciences more generally where the goals often include basic as well as applied research. Basic research refers to our interest in understanding the underpinnings of various phenomena—what, why, when, and how something happens. We may need to study the phenomenon under highly controlled conditions (e.g., nonhuman animal laboratory studies). Applied research refers to our interest in translating our knowledge toward goals of everyday life and in applied settings. For example, we want to understand as much as we can about stress and its impact on functioning and basic research has elaborated all sorts of features (e.g., how stress affects aging, the immune system, onset of depression) but we are also interested when possible to apply that information to alleviate stress (e.g., in everyday

life, for special groups who are exposed to harsh environments, war and trauma).

1.3.1: Definition and Its Components

Methodology refers to the diverse principles, procedures, and practices that govern scientific research. Methodology will be used as an overarching term that includes several distinguishable components, as noted in Table 1.1.

Table 1.1: Five Components of Methodology

Component	Definition
Research Design	Refers to the experimental arrangement or plan used to examine the question or hypotheses of interest. There are many designs, which we will cover and see how they work to help reach valid inferences.
Assessment	Refers to the systematic measures that will be used to provide the data. There are many different types of measures, multiple measures within each type, and more importantly for our purposes considerations to guide how to select measures.
Data Evaluation	Refers to the methods that will be used to handle the data to characterize the sample, to describe performance on the measures, and to draw inferences related to the hypotheses. You may recognize this as familiar statistical significance testing, but data evaluation is much more than that and even sometimes less (no statistical tests are used with some research designs).
Ethical Issues and Scientific Integrity	Refer to a variety of responsibilities that the investigator has in the conduct of the study and can encompass all of the other components of methodology (e.g., design, data evaluation, and communication of findings). Ethical responsibilities are to research participants (e.g., their rights and protections) and adherence to professional standards of one's discipline (e.g., ethical codes). Scientific integrity includes responsibilities to the scientific community (e.g., transparency, accurately reporting findings) and also is part of professional standards and ethical codes.
Communication of Findings	Refers to how the findings will be communicated to others in many different venues (e.g., journal articles of empirical studies, review articles) including the media (dissemination of information to the public via TV, radio, and the Web). There are many issues that emerge related to core issues of science (e.g., transparency of methods), but also challenges as what and how we communicate might be very different for colleagues and for the press.

1.3.2: Using Methodology to Answer Critical Questions

We will take up each of these aspects of methodology and present them separately to ensure each is given its fair treatment. As a reader, you may be especially interested in learning the concrete facets of methodology to answer critical questions to conduct a study, such as:

- How do I select a research question?
- What participants or subjects should I use?
- How do I decide exactly what measures to include in the study?[1]

We will certainly address specific practices and procedures to be of help. Yet, it is critical to consider broader issues underlying those practices and guiding principles. The broader issues are not some academic challenge with little impact. Just the opposite, once the overarching principles or reasons for various practices are understood, investigators—you and me—often have more flexibility in selecting concrete practices for our study.

Consider, for example, random assignment of participants to experimental conditions in a study. All the participants come to the study and are assigned in random order to groups (e.g., group 1 receives some task to induce happiness; group 2 receives some task to a neutral or slightly negative emotion). Random assignment is a core tenet of experimentation. The practice of random assignment, i.e., how exactly one does that is important and covered later.

> Yet, why do we do random assignment, and does it serve the goal we have in mind? We will discuss that too, and once we do it is easier to see that random assignment is not always critical, not problem free, and often goals to which random assignment is directed can be served in other ways.

This is not a text taking positions on key practices like random assignment; it is a text designed to develop black-belt methodologists and as part to that to equip you with a wide range of methods to solve and address the questions of interest to you. When one designs a study or reads a study that has been completed by others, knowledge about the practices and procedures is important. Yet the principles and rationales underlying those practices are critically important as well.

1.4: A Way of Thinking and Problem Solving

1.4 Analyze some of the key concepts that guide scientific thinking and problem solving

Methodology refers to a *way of thinking* and problem solving, in addition to the more concrete features we will discuss later in the text. That way of thinking is how we approach understanding the world around us. There are guides we follow, and these are worth noting and illustrating here before we address them in greater detail later in the text.

1.4.1: The Role of Theory

In science we want to explain what things are, how they work, how they relate to other phenomena, how they come about, and so on.

Theory at the most general level refers to an explanation.

That is, what phenomena and variables relate to each other, how are they connected, and what implications can we draw from that? We want to describe, predict, and explain, and theory can tie this all together. It is helpful to distinguish the findings that are obtained in a study from the conclusions the investigator may reach. The distinction is important for understanding theory as well as methodology.

1.4.2: Findings and Conclusions

> The *findings* of a study refer to the results that are obtained.

This is the descriptive feature of the study or what was found. A statement of a finding might be that one group was better or worse than another.

> The *conclusions* refer to the explanation of the basis of the finding, and this is the interpretative and theory part.

For example, as a sample finding, we know that corporal punishment of a child in moderate-to-severe doses (more than once per week, used as a primary discipline, not injurious physically and not necessarily at the level of physical abuse) is related to (correlated with) greater aggression on the part of the child. Children who are physically hit a lot as part of their punishment at home tend to be much more aggressive at school (more fighting, bullying). That is the finding—merely descriptive and factual—even though it may not mean for all children, in all families, and in all cultures and countries.

As for conclusions, we now would like an explanation of why corporal punishment and aggression are related. But we do not need some casual explanation from everyday life (e.g., "The kids are rotten and need to know their place and if anything punishment probably tames them!). We need a little more, to say the least. Specifically, we want theory that explains the relation and allows us to generate hypotheses that will guide us to elaborate on the explanation, to test the theory, and to revise and expand as needed.

> *Why a theory?* Well, we want to understand in part to learn some of the roots of and paths to aggression and also possibly to intervene or to prevent aggression. It is too quick to just say, "stop hitting your kids and they will not be aggressive," even though there are many reasons we would like parents to stop hitting their children.

Among the explanations, maybe children who are more aggressive lead their parents to extremes of punishment. Instead of nagging, reprimands, and shouting, the parents eventually escalate in an effort to stop seemingly uncontrollable aggressive behavior. This theory suggests that aggression in the child may have actually caused aggression in the parent. Alternatively, since so many things (e.g., aggression, depression, suicide, low key temperament, sense of humor, conscientiousness, love of methodology) run in families, perhaps the parents' aggression and the child's aggression do not influence each other very much at all. Rather, maybe they share common genetic origin and aggressive behavior in the parent and child reflects that. We could generate more explanations, but the goal is not merely to generate explanations but to move to empirical tests of one or two that we have identified. In passing it is useful to note that three explanations: parent modeling of aggression leads to more aggression in the children, child behavior and provoke parent aggression, and that there are shared genetic influences all have some support but the first explanation appears to be the stronger influence (see Moffitt, 2005).

We generate explanations to draw implications. Those implications are hypotheses that elaborate what might be going on and help us move forward.

> *If exposure to parental aggression leads to aggressive behavior in the child, how could we ever test that?* Among the options, bring young children in the laboratory and have some children watch movies or video clips of aggressive behavior and other children watch movies or clips of social interaction that are not aggressive. Then give the children the opportunity to show aggression (e.g., in relation to a doll or press one of two responses indicating what they would in a particular situation presented on a video—hit the other person or walk away).

This is merely one little test of whether exposure in principle can increase aggression, even if temporary and restricted to a lab setting. Let us not get too far into the example and lose the larger point. We select an explanation that accounts for (ties together, connects) our original facts (findings) and use that explanation to obtain more findings. In the process, we revise our theory to account for new facts including predictions that were supported or not supported. In the end, we want as full an explanation as possible. I am simplifying but will elaborate a bit in an example below.

1.4.3: Additional Information Regarding Findings and Conclusions

In everyday life, "theory" sometimes emerges with a different meaning. If someone says, "Oh, that's just a theory" or that is "theoretical" that meaning often refers to something that is pure speculation, hardly proven, and just a tale. This emerges in the ongoing debates of "creationism" and "evolution." As an explanation of how human and nonhuman animals emerged, there are many weighty issues in that debate including different ways of knowing (by faith, by science). Even so, among the many issues is a different use and meaning of the word "theory." When scientists use that term "evolution" is not a "theory" in a speculative sense. Rather it is an explanation developed with data from multiple sciences

(e.g., fossil record from geology, tracking development within and among from molecular and genetic measures, and viewing evolutionary processes actually unfold in the lab [studies of thousands of generations of yeast] spanning decades).

Evolution explains these facts and makes useful predictions, many supported by further facts, and so on. Creationists would not be expected to use that notion of theory, but are more apt to say, this is speculative and not proven. That view is not simply wrong at all. Much in evolution as scientists use that term is NOT proven or clear. All the mechanisms through which species change are not known (but some are), and there is much speculation about how we got from there (first day earth counted as a planet) to here (billions of years later with millions of plant and animal species and music groups with the weirdest names). No theory explains all of that, so there is indeed speculation involved. Yet, we know a lot and can even monitor and alter "evolution" (change and adaptation of bacteria, for example, to watch evolutionary change in response to environmental forces) in a laboratory (e.g., Wiser, Ribeck, & Lenski, 2013). As a way to explain scores of findings, evolution as a theory is on solid ground that is not speculative. Yet, this does not directly address the full range of concerns and points of creationists.

For this text, for evaluating research, and for your possible professional careers in any of the sciences, theory is that explanation or model we develop to guide our next steps in science. We want to explain and understand, and merely piling up facts and correlations will not do that at all. So we know that depression increases the risk for heart attack and that heart attack increases the risk for depression, and that if one has a heart attack and depression they are at much greater risk (than if they had just one of those) of dying (e.g., Lichtman et al., 2008). My God, these "facts" or the findings scream out for understanding.

> *What could be going on here that explains these relations?* One theory might focus on diet. Perhaps depressed individuals have lard omelets, fried chicken nuggets, and chocolate cheese cake (just a little sliver or two) for breakfast each morning and that diet increases the likelihood of heart attack. Well, that could be tested easily.

We might do a survey of individuals matched in age, sex, and education, but who vary in depression, and ask about what they eat. But as explanations go, it already looks weak because it does not explain the other direction, heart attack leading to depression, unless you believe the same diet would lead to heart attack patients becoming morose. That is not likely, but you may have a good explanation (theory) for that. Findings often are intriguing and raise a puzzle to solve. Theory helps generate the ideas for research; methodology includes the strategies to help us obtain the answers.

1.4.4: Parsimony

As we select our theory or explanation, we are guided by parsimony as a critical concept and way of thinking in science. Parsimony is not that cute little curly green vegetable that almost no one eats and is used to garnish the main course when restaurants bring you your food. Rather, parsimony is an accepted principle or heuristic in science that guides our interpretations of data and phenomena of interest.

> *Parsimony refers to the practice of providing the simplest version or account of the data among alternatives that are available.*

This does not in any way mean that explanations are simple. Rather, this refers to the practice of not adding all sorts of complex constructs, views, relationships among variables, and explanations if an equally plausible account can be provided that is simpler. We add complexity to our explanations as needed. If there are two or more competing views that explain why individuals behave in a particular way, we adopt the simpler of the two until the more complex one is shown to be superior in some way.

Apart from parsimony, there are other names for the guideline and they convey the intended thrust. Among the other terms are:

- The principle of economy
- Principle of unnecessary plurality
- Principle of simplicity
- Occam's razor

Where was the name "Occam's razor" derived from?

The term emerged from William of Ockham (ca. 1285–1349), an English philosopher and Franciscan monk. He applied the notion that makes this principle sound more complex; he proposed that plurality (of concepts) should not be posited without necessity in the context. That is, he believed that we ought not to add more concepts (plurality) if they are not needed to explain a given phenomenon. Supposedly, his frequent and sharp invocation of the principle accounts for why the term "razor" was added to his (Latinized) name to form Occam's razor.

1.4.5: How Parsimony Relates to Methodology

Parsimony relates to methodology in concrete ways. When an investigation is completed, we ask how to explain the findings or lack of findings. New concepts and more complex concepts may be used than existing concepts that are simpler, already available, and useful in describing many findings beyond those of the investigator. The investigator

may have all sorts of explanations of why the results came out the way they did. Methodology has a whole set of explanations that may be as or more parsimonious than the one the investigator promotes. Before we look to any new or complex explanation, we reach into our basket of already available explanations from every day as well as from prior scientific knowledge and ask ourselves, "Is there anything in the basket that can explain the data without adding more or more complex explanations?" For example, sightings of unidentified flying objects (UFOs) raise parsimony in the following way. We know that many concepts that are currently available explain the sightings that many people report. Meteorites across the sky (so-called "shooting stars"), odd patterns of temperature inversion in the sky, and military tests of secret equipment are among three parsimonious explanations and actually can account for many sightings. Indeed, one of these alone can explain many different sightings. So the question of parsimony here—can these simpler and well-established explanations be used? We only go to one that is more complex if they cannot.

> Is science against the notion of UFOs, or are scientists anti-flying saucers? Not at all, and indeed science is open to flying cups and saucers. For or against is not the issue.

Parsimony is a point of departure—can we explain something with concepts we have and without adding new complexities. In the case of UFOs, perhaps there are many sightings not explained by these existing concepts, and we have to go to other interpretations and creep slowly to add complexity a little at a time and as needed. We do not immediately jump to the idea of green Martians with hostile intent who have to gather minerals and food (humans) because they did not manage climate change on their planet very well. Way too many concepts here—always begin—what is the most parsimonious explanation we need to account for what we know, what the data show, what the facts are.

So let us say, we have a smartphone photo of what looks like an object in the sky. It is likely one of the explanations I already mentioned will be parsimonious—let us say for the moment we consider the photo to be of a meteor. Now new data come in. Say, we have in addition to a citing of something in the sky, now remnants of a "space ship" made out of materials very rare on earth and with a "map" inside that is in a never-before-seen set of symbols (language). With additional data, parsimony still argues for simplicity, but a meteor citing in the sky cannot explain the data (findings). Now we move to something more complex, which might be a hoax, visitors from a non-earthly place, or the equivalent. Parsimony requires accounting for what we find but simply.

A well-known illustration of competing interpretations is from cosmology and pertains to the orbiting of planets in our solar system. Nicolas Copernicus, a Polish scientist and astronomer (1473–1543), advanced the view that the planets orbited around the sun (heliocentric view) rather than around the earth (geocentric view). This latter view had been advanced by Claudius Ptolemy (ca. 85–165), a Greek astronomer and mathematician. Ptolemy's view had dominated for hundreds of years. The superiority of Copernicus's view was not determined by public opinion surveys or the fact Ptolemy was no longer alive to defend his position. Rather, the account could better explain the orbits of the planets and the varying brightness of planets and stars and did so more simply with fewer explanatory concepts. This is a case of parsimony or simplicity between the views but also more than parsimony because the Copernicus view could explain some of the data in a much better, cohesive way.

1.4.6: Plausible Rival Hypothesis

Plausible rival hypothesis is another key concept that guides scientific thinking (Campbell & Stanley, 1963; Cook & Campbell, 1979). Think of this concept as a methodological sister of parsimony; both concepts relate to interpretation of findings, and both represent critical features of thinking methodologically.

> *A plausible rival hypothesis refers to an interpretation of the results of an investigation on the basis of some other influence than the one the investigator has studied or wishes to discuss.*
>
> The question to ask at the completion of a study is whether there are other interpretations that can plausibly explain the findings. This sounds so much like parsimony that the distinction is worth making explicit.

Table 1.2: Distinction between Parsimony and Plausible Rival Hypothesis

Parsimony	Plausible Rival Hypothesis
Parsimony refers to adopting the simpler of two or more explanations that account equally well for the data.	This hypothesis has a slightly different thrust. At the end of the investigation, are there other plausible interpretations we can make of the finding than one advanced by the investigator?
The concept is quite useful in reducing the number and complexity of concepts that are added to explain a particular finding.	Simplicity of the interpretation (parsimony) may or may not be relevant.
Parsimony is about the minimum of ideas or concepts we need to explain what we have observed.	At the end of the study, there could be 2 or 10 equally complex interpretations of the results, so parsimony is not the issue.

1.4.7: An Example of Plausible Rival Hypothesis

For example, a new diet guru suggests that multi-berry fruit bars two times per day will increase one's intelligence quotient (IQ) and self-reported quality of life. To test that,

an investigator might recruit 20 volunteers and evaluates their IQ and quality of life before the diet begins. After initial testing, each participant gets a supply of fruit bars and downloads a fruit-bar reminder "app" (application). Twice a day, each participant receives a fruit bar text message and replies if a bar was eaten. After a month of the fruit bars, all participants return and get tested again. Sure enough, the findings show that IQ and quality of life increased—amazing. Now our investigator discusses how the fruit bars work and how they could change our lives.

- Are there any plausible rival hypotheses that might explain the effect that our investigator attributes to the fruit bars? Yes, one of these is called *testing*. As it turns out, individuals often improve on a measure (e.g., intelligence, personality, symptoms of psychopathology) when they are re-tested. Not always but often. So one rival hypothesis is the effect could be due to repeated testing, and the same results would have occurred if the group did not eat the fruit bars or only ate the wrappers of the bars.
- Is retesting really plausible? Yes, that is an area of research we already know about. This one-group study needs a second group at least that had the first and the second testing but with no fruit bars or some placebo bar! That group, if it did not change, makes testing no longer a plausible rival hypothesis or if the groups changed in the same way (no differences between groups) then testing may be a plausible explanation for the changes in both groups.

I hasten to add that plausible rival hypotheses can be parsimonious, so the concepts overlap. In the above example, the plausible rival hypothesis is repeated testing. Testing effect versus fruit-bar effect are two interpretations. For this study, both may be plausible and perhaps equally plausible. Parsimony helps because testing can explain findings from many studies and across situations in which repeated tests are provided. Thus, beyond this one study, parsimony has the advantage of one concept (testing) that explains many findings. We do not need fruit bars as an explanation until we rule out testing. Plausible rival hypotheses still can be distinguished because there are many explanations beyond testing that might explain the finding.

Methodology is all about the conclusions that can be reached from a study and making one interpretation of the findings more likely (plausible) than other interpretations.

> *How does one identify plausible rival hypotheses?* Well, many of them are well codified, and it is important to know exactly what they are before proceeding with one's own study and then when evaluating the studies of others.

The next chapters will provide the main rival explanations, and these too constitute the critical steps to methodological thinking. Methodological thinking includes a certain type of inquiry and skepticism insofar as it is fine, even better than fine, to ask there other plausible interpretations or explanations than the one that is being promoted. This is not just a skepticism one direct only toward others; we direct it to our own studies to optimize the clarity of the conclusions we reach.

1.5: The Semmelweis Illustration of Problem Solving

1.5 Discuss the importance of Semmelweis's usage of a scientific way of thinking to solve a problem.

Developing explanations (theory) and testing theory by generating hypotheses, adhering to parsimony, and considering plausible rival hypotheses are way too abstract to convey how they are used or that they really make a difference to anyone. Science as a way of thinking and drawing on these concepts is nicely illustrated by the story of Dr. Ignaz Semmelweis (1818–1865), a physician who worked at the Vienna General Hospital in Austria.

1.5.1: Illustration: Saving Mothers from Dying

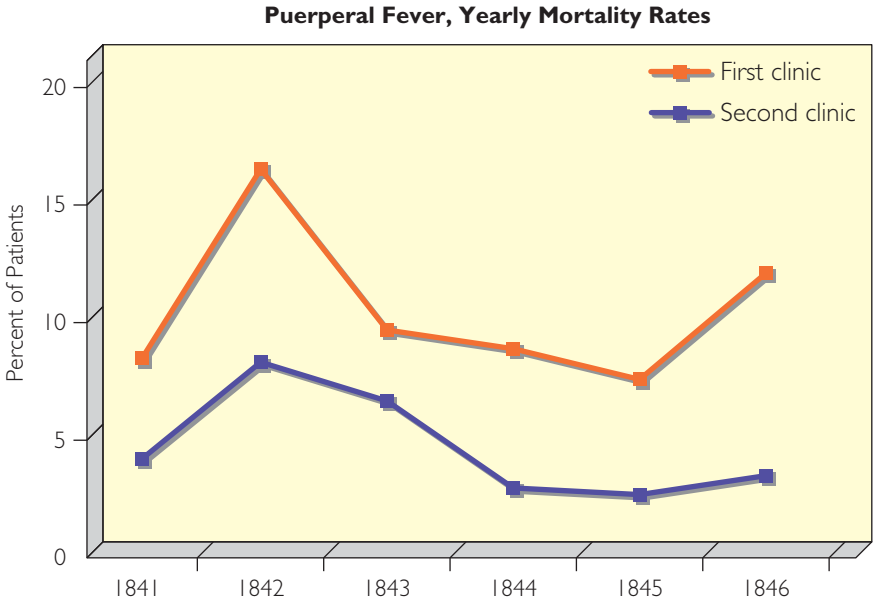
Vienna General was a large hospital used for medical training for doctors throughout Europe in part because of the availability of many cadavers that could be used for study. Semmelweis worked in obstetrics and was involved in examining patients, supervising difficult deliveries, and teaching students.

At this one hospital, there were two separate clinics for delivering babies. Women were admitted to the clinics on alternate days as they arrived to deliver their babies. The first clinic was used as a teaching service for medical students. The second clinic was used for instructing midwives only. Both clinics delivered babies, and there were no differences in that regard. One difference between the clinics was well known at the hospital and also by prospective mothers. The rate of mothers dying while at the first clinic was high; 10–18% of the mothers died from a disease while in the hospital. The rate of mothers dying while in the second clinic was much lower at about 4%. As shown in Figure 1.1, over a period of years the differences between the two clinics were consistent and dramatic.

Women coming to the hospital knew of this and begged not to be admitted to the first clinic. In fact, many women "pretended" to be on their way to the hospital but delivered their babies in the street (called street births) just to avoid the first clinic. (They would still qualify for state

Figure 1.1: Mortality Rates for the Two Clinics at the Vienna Hospital

Higher rates of death for the first clinic (top line) from 1841 to 1846.



child care benefits if they were on the way to the hospital.) The disease from which the mothers died while in the hospital was puerperal fever (also known as childbed fever), which is a form of septicemia or sepsis.[2]

1.5.2: Additional Information Regarding the Semmelweis Illustration

Semmelweis wanted to explain (theory) why the death rates were so different between the two clinics. Add to the complexity, the street-birth mothers who delivered their babies under less desirable conditions rarely died of the disease.

What was so special about the first clinic?

He ruled out differences in the first and the second clinic related to crowding—indeed the clinic with fewer deaths was more crowded. There were no differences in religious practices among the patients that might somehow influence healing. Also, it is not plausible to believe that the mothers at the different clinics were different types of people in some way. Assignments were made to the clinic every other day—not exactly random but still no basis for any systematic bias that could explain the different death rates. The main difference was that one clinic trained medical students and the other did not. But that is a description of the differences between the clinic and still not an explanation of mortality rates.

A tragedy happened while Semmelweis was briefly out of the country. A senior physician and colleague of his at the hospital became ill. That doctor was conducting autopsies as part of training of medical students. During one of these autopsies, one of the students accidentally pricked the finger of the doctor with the scalpel used in the autopsy. Very shortly thereafter, the physician became very ill with a massive infection throughout his body (lungs, membranes of the heart, and brain) and died. Semmelweis learned of his colleague's death and immediately returned to the hospital. He could see from autopsy that his colleague had died of the disease identical to those contracted by the mothers. Now he developed a theory, i.e., a possible explanation to account for the facts. The facts now included the higher death rate of the first clinic and the death of his colleague at that clinic, following a wound of a scalpel used during an autopsy.

He reasoned that there must be "cadaverous particles" (something from the cadavers) that were passed from the scalpel to his colleague and also perhaps to other mothers (because instruments were not cleaned nor was it routine to wash hands between seeing patients). These particles caused the disease—that was his theory at least.

1. The first challenge of the theory: could the theory explain why many deaths were at the first clinic but fewer at the second clinic? Yes—at the second clinic, no autopsies were done and the midwives were not trained in that. Thus, there was no spread of the disease from doctors doing autopsies to patients from equipment or from their hands.
2. A second challenge for the theory was to test the hypotheses that might follow. If there were particles (think bacteria, germs) on instruments and even the hands of the doctors, try to get rid of them (the germs, not the doctors). Getting rid of the supposed particles would reduce the death rate, or at least that would be predicted from the theory.

1.5.3: A New Procedure

Semmelweis started a new procedure of using a chlorinated solution (with a compound used in bleach) to have doctors wash their hands between autopsy work and examining patients. The solution was used because he found it to work on removing the smell of infected autopsy tissue and perhaps that fluid would destroy whatever material might be transmitting the infection to the patients.

When the washing procedure was implemented, mortality rated dropped 10-fold. The death rate of moms went from 18.3 (in 1847) before the hand washing was started to under 2% in the months after and down to 0 (see Figure 1.2).

He extended his washing procedures to include all instruments that would make contact with patients (e.g., in labor) and continued to show that puerperal fever was virtually eliminated from the ward.

For a variety of reasons, Semmelweis's views that deaths could be traced to the lack of cleanliness were ignored or rejected. Well-known doctors at the time publicly denounced his views. Semmelweis's breakthrough was before Louis Pasteur developed the "germ theory," i.e., that there were active bacteria that might be passed along. Also views of disease at the time emphasized an imbalance of body humors (yellow bile, black bile, phlegm, blood), a carryover from Greek medicine, bad air from atmospheric and cosmic influences (e.g., influences of planets and stars), and the idea that diseases at the clinics were simply contagious and perhaps all caught something on the ward at the first clinic. This was not helped by data from autopsies. Sepsis symptoms were not identical in all cases, so autopsies did not show uniform disease processes. This could lead one to a view that all individuals died of different or a few separate diseases—not very parsimonious but still possible and to some even plausible.

Semmelweis's view was plausible. It was also parsimonious—a single interpretation could explain:

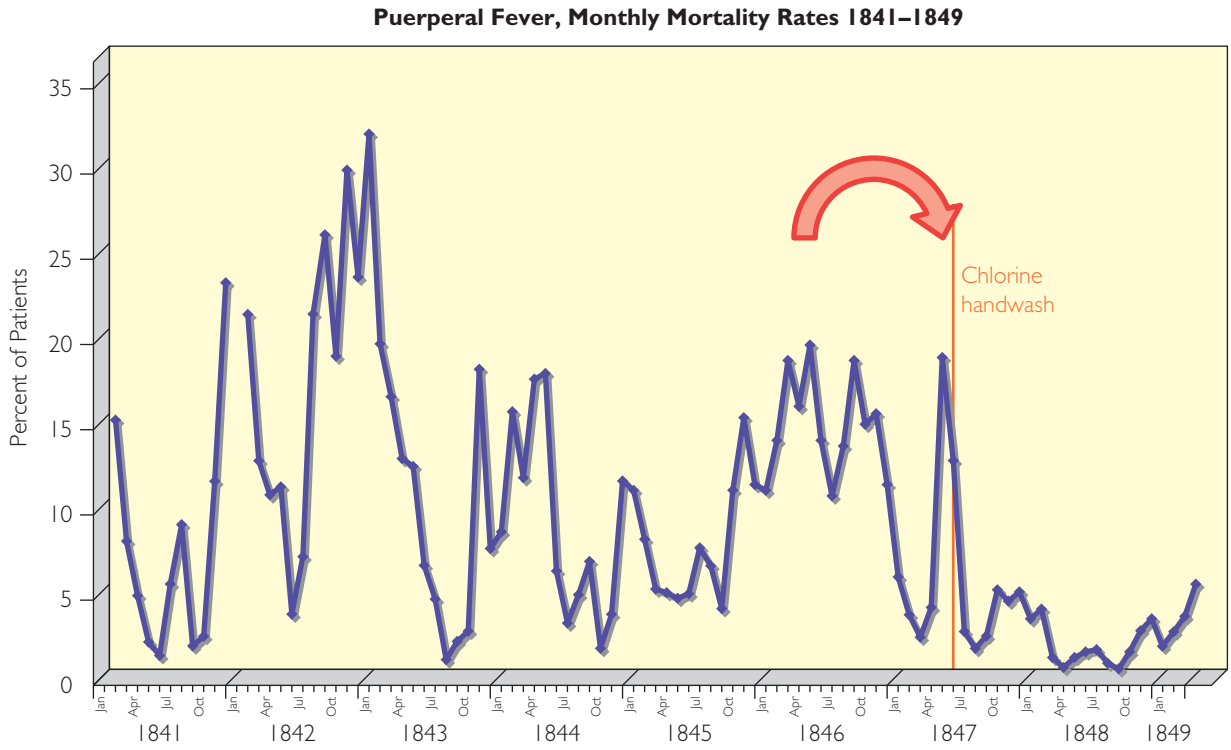
- The deaths of many people at the first clinic
- The death of his colleague
- The differences in death rates between the first and the second clinic
- The fact that street birth moms, delivering under unsanitary conditions, did not show the high disease rates
- The reduction of deaths by testing his cadaverous particle theory and by cleansing procedures designed to disinfect the cadaverous particles

There are many ways in which this is science at its best—developing a parsimonious and plausible explanation theory to account for diverse facts and testing the hypothesis that follows from that theory.

Over the years, Semmelweis took jobs at different hospitals and there was a pattern that emerged. When he added his cleansing procedure at a new hospital, death rates of mothers at that hospital greatly declined. In current work, this would be called replication, i.e., repetition of results using the same intervention but with different patients and

Figure 1.2: Mortality Rates for the First Clinic

Reduced death rate once hand washing begun (see the curved arrow and vertical line during the year 1847).



at different hospitals. The original effects obtained at the Vienna hospital are not very likely to be a fluke given repeated replication. The most plausible hypothesis is that the intervention somehow eradicated the particles. Also replicated (sadly), when he left a hospital and took a new job, the previous hospital tended to revert back to their old noncleanliness procedure and death rates increased. Thus, the higher death rate also was replicated when the procedure was abandoned. In the ensuing years after Semmelweis's breakthrough in medical care and the scientific basis for his work became clear—germ theory, I mentioned work of others on cleanliness in surgical treatment. What was once completely rejected now became standard clinical practice. In his lifetime, Semmelweis was demeaned rather than recognized but this all changed in time.[3]

1.5.4: General Comments

Semmelweis's contributions to alter medical practices stand on their own, but the story is noted here because of his use of a scientific way of thinking to solve a problem. He might not have thought of it quite that way; indeed, long before the formalization of current methodological practices, there are many examples where a problem was addressed by trying to understand, experimenting with possible solutions, and seeing if a solution once demonstrated could be repeated.

Many relations in science are more complex than the one in the Semmelweis story. Find a pathogen, try a cure, and have an effect. Also, many relations that exist in nature cannot be so easily discerned, and many that are easily discerned are not genuine relations (as noted with the confirmatory bias). We draw on science and the methods of science to help sort things out. For example, I mentioned previously the relation of heart disease and depression and how one increases the risk for the other and that having both increases the risk of death. Several studies were needed to rule out various explanations for these relations, and in those studies many practices and procedures were followed to minimize bias.

Methodological practices are intended to help reach conclusions with minimum ambiguity and bias. At the completion of a study, the explanation one wishes to provide ought to be the most plausible interpretation. This is achieved not by arguing persuasively, but rather by designing the study in such a way that other explanations do not seem very plausible or parsimonious. The more well designed the experiment, the fewer the alternative plausible explanations that can be advanced to account for the findings. Ideally, only the effects of the independent variable could be advanced as the basis for the results.

Critical Thinking Questions

1. What are some of the limitations of human perception and cognition that interfere with acquiring knowledge about the world?
2. What is theory, and why is it important as part of research?
3. Parsimony and plausible rival hypothesis are so key to science. What do these concepts mean?

Chapter 1 Quiz: Introduction

Chapter 2
Internal and External Validity

Learning Objectives

2.1 Report four types of experimental validity used to evaluate the methodology of a study

2.2 Define internal validity

2.3 Classify some of the different threats to internal validity

2.4 Report how instrumentation serves as a threat to internal validity

2.5 Summarize each of the additional threats to internal validity

2.6 Review the four main circumstances of potential threats to internal validity

2.7 Discuss the importance of determining the relevance of a threat to internal validity in order to manage it

2.8 Define external validity

2.9 Summarize different threats to external validity

2.10 Classify each of the additional threats to external validity

2.11 Evaluate the idea of proof of concept

2.12 Examine the importance of determining the relevance of a threat to external validity before it is managed

2.13 Analyze the similarities and differences between internal validity and external validity

The important concept of **plausible rival hypothesis** addresses those competing interpretations that might be posed to explain the findings of a particular study. Methodology helps rule out or at least make implausible competing interpretations. An experiment does not necessarily rule out all possible explanations. The extent to which it is successful in ruling out alternative explanations is a matter of degree. From a methodological standpoint, the better the design of an investigation, the more implausible it makes competing explanations of the results. There are a number of specific concepts that reflect many of the interpretations that can interfere with and explain the results of a study. The concepts are critical too as they serve as a methodological checklist so to speak. When planning a study or evaluating the results of a completed study, it is extremely useful to know the many concepts we cover and how they will be or were handled in the design of the study. A fascinating study, with all sorts of niceties, and seemingly exemplary features can be undermined by not addressing the concepts we discuss.

2.1: Types of Validity

2.1 Report four types of experimental validity used to evaluate the methodology of a study

The purpose of research is to reach well-founded (i.e., valid) conclusions about the effects of a given experimental manipulation or intervention. Four categories of types of experimental validity organize the many concepts used to evaluate the methodology of a study. These include:

- Internal
- External
- Construct
- Data-evaluation validity[1]

These types of validity serve as a useful way to convey several key facets of research and the rationale for many methodological practices, and as well to remember types of problems that can emerge in designing and interpreting a study. Table 2.1 lists each type of validity and the broad question each addresses. Each type of validity is pivotal. Together they convey many of the considerations that investigators have before them when they design an experiment.

Table 2.1: Types of Experimental Validity and the Questions They Address

Type of Validity	Question or Issue
Internal Validity	To what extent can the intervention rather than extraneous influences be considered to account for the results, changes, or differences among conditions (e.g., baseline, intervention)?
External Validity	To what extent can the results be generalized or extended to people, settings, times, measures/outcomes, and characteristics other than those included in this particular demonstration?
Construct Validity	Given that the experimental manipulation or intervention was responsible for change, what specific aspect of the manipulation was the mechanism, process, or causal agent? What is the conceptual basis (construct) underlying the effect?
Data-Evaluation Validity	To what extent is a relation shown, demonstrated, or evident between the experimental manipulation or intervention and the outcome? What about the data and methods used for evaluation that could mislead or obscure demonstrating or failing to demonstrate an experimental effect?

> Methodology is a way of thinking that relates directly to how one thinks about a study.

Once these validities have been mastered, they form a way of thinking about and evaluating any scientific investigation, not just in psychological science. The methodological strengths and limitations of any study are encompassed by these considerations, and if one knows the various types of validity and what can go wrong one can evaluate any study very well. Apart from how to think about a study, the problems that can emerge with each type of validity translate to specific methodological practices to improve the quality of an investigation at the design stage.[2]

> In designing an experiment, it is critical for investigators to identify their purposes and specific questions quite clearly at the outset. The reason is that it is impossible to design and execute an experiment that addresses each type of validity perfectly.

Occasionally, a decision to maximize one type of validity can be at the expense of others—trade-offs so to speak. Investigators prioritize types of validity and manage potential problems to ensure that their hypotheses are well tested. This chapter discusses internal and external validity. These types are presented first because they are relatively straightforward and reflect fundamental concerns.

2.2: Internal Validity

2.2 Define internal validity

To help orient us, consider two basic and likely familiar terms, independent and dependent variables:

- Independent variable of course is the experimental manipulation or variable we are evaluating to see if it has an effect.
- Dependent variable is the outcome or measure we are examining to reflect the impact or effects of the independent variable.

In any study, we wish to draw conclusions about the effects of the independent variable. All this sounds so straightforward and simple, and this is a good place to start.

An investigation cannot determine with complete certainty that the independent variable accounted for change. However, if the study is carefully designed, the likelihood that the independent variable accounts for the results is high or very plausible. When the results can be attributed with little or no ambiguity to the effects of the independent variable, the experiment is said to be internally valid.

> Internal validity refers to the extent to which an investigation rules out or makes implausible alternative explanations of the results.

Factors or influences other than the independent variable that could explain the results are called threats to internal validity. That is, these influences the investigator may not have carefully considered "threaten" or jeopardize the interpretation the investigator wishes to make about the results. The demonstration becomes unclear because these factors were not controlled. Threats to validity are the problems that could emerge in the design or execution of the study. In terms of what to know in designing your own studies and evaluating others, mastery of each threat is really helpful.

2.3: Threats to Internal Validity

2.3 Classify some of the different threats to internal validity

Several threats to internal validity have been delineated long ago (e.g., Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). They remain relevant and, as one can

Table 2.2: Major Threats to Internal Validity

Specific Threat	What It Includes
History	Any event (other than the intervention) occurring at the time of the experiment that could influence the results or account for the pattern of data otherwise attributed to the experimental manipulation. Historical events might include family crises; change in job, teacher, or spouse; power blackouts; or any other events.
Maturation	Any change over time that may result from processes within the subject. Such processes may include growing older, stronger, healthier, smarter, and more tired or bored.
Testing	Any change that may be attributed to the effects of repeated assessment. Testing constitutes an experience that, depending on the measure, may lead to systematic changes in performance.
Instrumentation	Any change that takes place in the measuring instrument or assessment procedure over time. Such changes may result from the use of human observers whose judgments about the client or criteria for scoring behavior may change over time.
Statistical Regression	Any change from one assessment occasion to another that might be due to a reversion of scores toward the mean. If clients score at the extremes on one assessment occasion, their scores may change in the direction toward the mean on a second testing.
Selection Biases	Systematic differences between groups before any experimental manipulation or intervention. Any differences between groups (e.g., experimental and control) may be due to the differences that were already evident before they were exposed to the different conditions of the experiment.
Attrition	Loss of subjects over the course of an experiment that can change the composition of groups in a way that leads to selection biases. Attrition affects other types of experimental validity as well.
Diffusion of Treatment	Diffusion of treatment can occur when the intervention is inadvertently provided during times when it should not be (e.g., return to baseline conditions) or to persons who should not yet receive the intervention at a particular point. The effects of the intervention will be underestimated if it is unwittingly administered in intervention and nonintervention phases.

see from examples, can be identified in contemporary research where they have not been suitably controlled. The challenge is to design a study that makes these threats implausible as an interpretation of the results. To the extent that each threat is ruled out or made relatively implausible, the investigation is said to be internally valid. Table 2.2 summarizes major threats to internal validity to provide an easy reference, but each type is highlighted here.

2.3.1: History

History refers to any event, other than the independent variable, occurring in the experiment or outside of the experiment that may account for the results. History refers to the effects of events common to all subjects in their everyday lives (e.g., at home, school, or work). The influence of such historical events might alter performance and be mistaken for an effect resulting from the experimental manipulation or intervention. It is important to be able to distinguish the effect of events occurring in the lives of the subjects from the effect of the experimental manipulation.

> Events that happen to all of us all of the time and our individual histories and experiences are not what is meant by history as a threat to validity. Rather, this refers to any event that happens to virtually all of the subjects that might explain the findings of an experiment.

This is a systematic bias or experience that the subjects receive while in the experiment that could explain how they responded on the dependent measures.

For example, if the experiment takes a short or a long time (e.g., over a period of 2 days or several years), historical events, events in the news (e.g., a natural or "manmade" disaster at a national or local level), or some other common experience might explain changes in the subjects or contribute to the results. An episode of school violence in the national news and all the associated publicity might be relevant to an experiment and the reactions that subjects have to an experimental manipulation. It could be that subjects are more anxious, stressed, or distracted as a function of the event and that reaction might plausibly explain the changes the subjects show from an assessment before the manipulation (pretest) to the assessment after the manipulation (posttest). For historical factors, one looks for influences that could affect all or most of the subjects. Examples are provided in the discussion of the next threat (maturation), which often goes together with history as a threat to internal validity.

> Although history usually refers to events outside of the experiment, it may include events that take place during the experiment as well. When subjects are run in a group, unplanned events (e.g., power blackout, medical emergency of one of the participants, fire drill) may disrupt administration of the intervention and reduce or enhance the influence performance of the participants.

Insofar as such events provide plausible explanations of the results, they threaten the validity of the experiment. This latter point is critical to note. One cannot criticize an experiment by merely saying, "maybe history was a threat

to validity and could explain the results." More is needed to show that:

1. There was a historical event that occurred.
2. It is a plausible interpretation of the findings.

2.3.2: Maturation

Changes over time also may result from processes within the subjects.

> Maturation refers to processes within the participants that change over time and includes growing older, stronger, wiser, and more tired or bored.

As with history as a threat, maturation may arise as a competing explanation of the findings when there is more than one assessment occasion and where some change within the individual might be a plausible explanation of change on the measure from one occasion to the next.

> Maturation is only a problem if the design cannot separate the effects of maturational changes from the intervention. That is, are changes within the individual over time a plausible explanation of the findings? The time frame of the study (two sessions spaced days or weeks apart; longitudinal study with multiple sessions over a period of years) is relevant to invoking this threat. In general, maturation has greater impact over time.

Maturation and changes associated with it are familiar in everyday life. It is not true that "time heals all wounds," but the expression does capture the fact that processes associated with time often lead to change. If one is studying treatment of a common cold or childhood anxiety, it is likely that maturational processes alone will lead to some changes. If one is testing a new treatment, it is important that maturation can be ruled out. For example, the common cold is likely to get better over time without an intervention. Similarly, many sources of anxiety in children dissipate with time. Maturation is a threat only if it is possible that such changes might explain the changes that occur in a study.

History and maturation often, but not invariably, go together as threats to internal validity. In any given case, it may not be easy to determine whether historical events or maturational processes accounted for change. In an experiment, the investigator must rule out that these changes associated with passage of time, whatever their basis (history, maturation), can be distinguished from the changes associated with an intervention (e.g., an experimental manipulation in a lab study, some psychosocial intervention).[3]

Critical Thinking Questions

Describe and give examples of how history and maturation can act as threats to internal validity.

2.3.3: Testing

> Testing refers to the effects that taking a test one time may have on subsequent performance on the test.

In an investigation, pre- and post-manipulation assessments might be given to evaluate how much an individual changes from one occasion to the next. The investigator may wish to attribute the change to the intervening experimental manipulation. Yet, practice or familiarity with the test or measures themselves may influence performance at the second testing. That is, changes at the second testing might be due to an experimental manipulation or the effects of repeated testing.

As an example, testing effects were evident in a study designed to prevent sexual assault among military personnel (Rau et al., 2011). Sexual assault in the military is a significant problem that affects primarily but not exclusively women. The military has all sorts of programs to monitor, treat, prevent, and litigate to help reduce and eliminate the problems (United States Department of Defense, 2012). This study provided a multimedia (lecture, videos) intervention to increase knowledge, empathy, understanding of military rules and policies, perspectives of women, and so on to better educate males.

Military participants received the intervention or control (nonintervention) condition; for present purposes, the interesting feature is that some participants received a pretest and a posttest and others received just a posttest. The goal of this latter feature was to see if repeated testing had any effect all by itself or influenced responsiveness to the intervention. The results indicated a testing effect for both treatment and control participants. That is, those participants who completed the measure on two occasions (testing) were higher in empathy about rape and also rejected myths about rape to a greater extent than did men who took the posttest alone. In other words, repeated testing, whether one received the intervention or not, led participants to have significantly more empathic and informed views of sexual assault even if they were in the control (nonintervention) group. Simply completing the measures on two occasions led to higher scores.

Beyond this example, we have known for years that repeated testing can influence many domains of functioning. Even without any special intervention, performance on measures of adjustment and personality sometimes improve and become more positive (e.g., adjusted) on the second testing occasion. This can also occur in educational contexts in which repeated testing leads to improvements in scores among college students—whether or not they receive training in some curriculum (e.g., Pattar, Raybagkar, & Garg, 2012).

> Repeated testing does not always lead to improvement, and even when it does there are obvious limits so that endlessly retaking some measure will not lead to continued improvement.

(I learned this the hard way; repeatedly taking the same IQ test 19 times and still could not get my score in the >100 range.) Repeated testing does not work like that, and as we talk about assessment and what goes into a person's score the reasons will become clearer later in this chapter. Yet, there is the equivalence of "practice" effects for psychological tests, and these can be mistaken for the impact of an experimental manipulation if the study is poorly designed. For now, it is important to note that any study that shows a change from pretest to posttest ought to control for testing as a plausible rival explanation of the results. Was the change from pre to post due to the experimental manipulation or just repeated testing? A well-designed study removes that ambiguity by using a control condition with repeated testing but without the special experimental manipulation or intervention.

2.3.4: History, Maturation, and Testing Combined

History, maturation, and testing often go together as threats to internal validity. For example, a recent study treated individuals who met psychiatric criteria for anxiety disorder (Rathgeb-Fuetsch, Kempter, Feil, Pollmächer, & Schuld, 2011). A goal was to see if treatment (cognitive behavior therapy) would be effective with patients with and without another disorder in addition to their anxiety disorder. All patients received the treatment, and no control condition was used. Both groups of patients (anxiety only or anxiety plus another disorder) improved on measures (of symptoms, cognitions, and avoidance). Of course in such a study, the authors would like to conclude that treatment was effective and equally effective whether or not the individuals had another disorder in addition to anxiety.

In light of the design and results and unaddressed threats to internal validity, we have to insert caution and skepticism about the interpretation of the findings.

- Was treatment effective?
- Was treatment needed?

History, maturation, and testing at the very least cannot be ruled out. We might not be able to identify a plausible historical event, so let us draw on maturation and testing. Why even bring these threats up?

Because these influences are known to effect change in many demonstrations, and we do not need more explanation than these threats (parsimony). This is not to say that maturation and testing did cause the change but that the study cannot really comment on the impact of treatment without these being controlled. Just one of them (e.g., repeated testing) cannot be ruled out (plausible rival hypotheses), and we must say that the study provided no evidence that treatment led to change.

2.4: Instrumentation as a Threat to Internal Validity

2.4 Report how instrumentation serves as a threat to internal validity

> Instrumentation refers to changes in the measuring instrument or measurement procedures over time.

The most common situation would be where ratings are made by judges or oneself and somehow the standards or criteria for making those ratings change over time. Changes in the dependent variable over the course of a study may result from changes in scoring criteria, rather than changes in what is being rated. Some examples will clarify.

To begin, you may easily recognize instrumentation in the context of athletic competition. In many sports (e.g., think Olympics) such as gymnastics and figure skating, there are no "objective" scores like points on a scoreboard, time (e.g., as in races), distance (e.g., how far some object is thrown), or accuracy (e.g., hitting a bulls eye). Rather judges rate performance of individuals in a given event.

> Here judges and their ratings are the "measure," and instrumentation raises the question, "Is there any change in the measure from one occasion (athlete being rated) to the next?" Instrumentation would refer to any changes in the scoring criteria that judges might unwittingly invoke over time.

Assume for a moment that a very superb gymnast performs first on a given event and receives perfect ratings from all of the judges. Will the criteria or standards for making judgments be any different for the next person who is to be rated, or if the next person performs identically, will she or he receive all perfect ratings? It is conceivable that the standards for ratings change a little over time. (Friendly advice—the next time you are in national or international competition for events where judges make ratings, walk over and explain instrumentation to the judges in a constructive fashion before you perform to make sure that they do not unfairly change the standards. I think my failure to do so has been the reason why I never win medals.) From a methodological perspective, the recommendation for Olympic performance evaluations or any situation in which repeated judgments have to be made is rather clear. All performances of the individual competitors could be recorded; each judge could see the recordings of all performances in random order (or balanced order). Then changes from who went first or last would not be likely to reflect any bias due to subtle changes in the measuring tool (i.e., judgments of raters). There still might be changes in criteria for any given judge, but those would not differentially bias one person across all of the judges, because who went first or last was random in the video recordings presented to the judges.

2.4.1: Some Examples Involving Instrumentation

Consider a dramatic example where raters and instrumentation are involved. Incarcerated individuals for many crimes can come before parole boards who determine whether the prisoners will be granted parole. An evaluation of multiple parole decisions revealed that in the morning and immediately after a lunch break, the likelihood of being granted parole is much higher than at other times (Danziger, Levav, & Avnaim-Pesso, 2011). Indeed as hunger (or fatigue) increases and as lunch time approaches, the chances of being paroled decrease but bounce up again right after the lunch break. The same raters are involved, and the result cannot be explained by severity of the crimes or types of prisoners. Instrumentation in this case, systematic change in raters with the added finding that there is a pattern associated with the timing of the break. (It was not clear from the study whether the "break" was the issue or being fully fed was the issue, but that does not detract at all from the instrumentation issue.)

***The methodological lesson*:** Instrumentation can operate when raters are used as a basis for assessment; *life lesson*, next time you are up for parole, be sure you are one of the first in the morning or immediately after lunch.

Instrumentation can greatly affect substantive conclusions about changes over time in clinically and socially relevant domains. A dramatic example pertains to the incidence (rate of cases) of autism spectrum disorder (ASD). This is a group of developmental disabilities characterized by impairments in social interaction and communication and by restricted, repetitive, and stereotyped patterns of behavior. Typically, ASD symptoms emerge before age 3 years. Widely circulated in the news is the fact that the rates of ASDs in the population keep rising. This was once considered a rare disorder, but more recent data suggest that as many as 1 in 88 children is affected (1 in 54 if just boys are considered) (Centers for Disease Control and Prevention [CDC], 2012). Since 2002 this was a 78% increase; since 2006 a 23% increase to 2008 when the last of this large survey was conducted. In a more recent study, the rate was even higher at 1 in 38 children (Kim et al., 2011). What to make of the rapid increase in the disorder?

Instrumentation is involved. The ways in which cases are being identified (referred to as ascertainment) are much more thorough and comprehensive than it has been before and as well the definition of what counts as the disorder has changed. That definitional change has gone from autism or autistic disorder, defined narrowly, to a continuum of symptoms in varying degrees (ASD). In relation to the present discussion, the measurement tools and their definition have changed and contribute to the different results. Research suggests that instrumentation is part of the explanation of why there has been a sharp increase in the reported rates of ASDs. Also, use of the diagnosis has increased due to familiarity of the public with the symptoms and the availability of clinical services, some of which may require the diagnosis for admission. In addition, the general view is that in fact there is a "real" increase with ASDs as well. It is more difficult to separate the different methods of assessment from the actual increase, although both are considered to be responsible for the high rates now evident.

> Changes in definitions are part of instrumentation when rates of assessment are used and compared over time.

Thus, crime rates; drug use; rates of arrest to specific infractions; traffic tickets; poverty; and even ethnicity in a city, state, and country can change over time as a function of changing definitions, instruments used to assess them, or care in identifying cases. What this means is that when one sees changes over time in a measure, the first query to make is about the assessment devices and whether any change was made in the definition or criteria that were used.

2.4.2: Additional Information on Instrumentation

Instrumentation usually is not considered as a problem when standardized paper-and-pencil tests are administered or when automated devices (e.g., press a touch screen) are used to score a response. Yet, the conditions or context of administration can greatly influence the nature of the measure, holding all of the items constant. For example, one study wished to evaluate the extent to which students who had an addiction to the Internet also showed various psychiatric symptoms (Dong, Zhou, & Zhao, 2011). The term "addiction" is used to reflect excessive use and dependence that leads to impairment in other domains of functioning (e.g., school, work, interpersonal relations). In this study, incoming college students were tested on a questionnaire (Hopkins Symptom Checklist 90) at a university in China, as part of routine assessment of the mental health status of students. A year later, students then completed an online measure of addiction that asked about various symptoms. The goal was to identify those who became addicted to the Internet now but were not originally when they completed the measure. With a total now of 59 students, they completed the original questionnaire (symptom checklist) again but this time as part of a research project rather than routine university administration of a measure. For this latter administration of the same measure, informed consent was needed. The results showed differences on some symptom domains (increase) from one test occasion to next.

The goal was to show that individuals who became addicted to the Internet also had other symptoms as well (depression, anxiety). What can we say about the conclusions? From the standpoint of instrumentation, the quite different test conditions from time 1 to time 2 preclude their comparison. On one occasion, the measure was not part of a study; on the next occasion, the students had to sign consent. Yes testing effects (repeated performance on the test) are a problem as well, but even more so the changing of the contexts in which the measure was administered. This is an example of instrumentation in which the measure (exact items of the symptom checklist) did not change but the context as part of the measurement conditions did change. The authors drew conclusions about changes in symptoms, but a key threat remains quite plausible as an alternative explanation.

In other instances, the measuring devices, instruments, and scoring procedures may be the same, but other contextual influences may change. For example, casual remarks by the experimenter at the time of the test administration might affect the subject's response and effectively alter the nature of the test and how the responses are obtained. For example, in a laboratory experiment on the reduction of arousal, stress, and anxiety, the experimenter might well say, "I'll bet you're really relieved now that the film (story, task) is over. Please complete this measure again." These different instructional sets or attitudes on the part of the experimenter are part of the measurement procedures. Conceivably, the different instructions preceding the measure could alter the assessment in systematic ways and lead to the report of less anxiety. The reduction may result from assessment changes, rather than from the experimental manipulation or changes over time due to other influences (history, maturation).

In general, it is possible that the instrument can change and be the same at the same time, even though this sounds contradictory. It is possible that the items remain the same (i.e., absolutely no change in the instrument, wording, or ways in which the instrument is administered). For example, a standard paper-and-pencil inventory (e.g., Beck Depression Inventory) might be administered at the beginning and end of the experiment. Obviously, the items are objectively the same on paper from one occasion to the next. Yet, the items may have different meaning because of the social context of a given point in time.

2.4.3: Response Shift

Response shift refers to changes in a person's internal standards of measurement.

This includes the case of judges who rate athletic performance or prisoners up for parole, as I mentioned. But the phenomenon can be more general any time there might be a change (shift) in values, perspective, or criteria that lead to evaluation of the same or similar situations, behaviors, and states, in a different way (e.g., Howard, Mattacola, Howell, & Latterman, 2011; Schwartz & Sprangers, 2000). For example, in clinical psychology, one can readily envision response shift (instrumentation) in the context of psychotherapy. Individuals are tested on some measure or a set of measures, go through some form of psychotherapy, and are retested on the same measure. It could be that clients did not "really" change after treatment in the problem domain (e.g., anxiety, tics, body image, and even weight) but have altered their standards in defining what a problem is. The clients may see themselves as improved because they know now that relative to what they thought before or relative to other people, their problems are minor. This concern actually has a label and is referred to as the *hello-good-bye effect* (Hill, Chui, & Baumann, 2013; Streiner & Norman, 2008). The term is based on the view that before treatment clients may have seen their lives as especially bleak or perhaps even distorted a little to get into treatment. At the end of treatment, they respond now by having altered their threshold for noting symptoms or seeing their lives differently even though the symptoms may not have changed. Here the actual functioning of the client and the measure itself (e.g., items, format) has not changed, but the standards for rating one's own functioning may have changed. Response shift reflects a change in *threshold* for answering a particular way. The threshold may be influenced by historical and maturational changes in the individual or the context (e.g., after treatment) in which the instrument is embedded.

Instrumentation as a threat to internal validity is any instance in which differences (e.g., between one assessment occasion and the next) might be attributed to a change in the instrument or to a change in the criteria (response shift) used to complete that instrument. It is important not to confuse testing (a threat addressed previously) with instrumentation. Both include measurement, and both relate to changes from one occasion to the next.

Testing refers to changes in the individual over time (due to experience and practice on the measure).

Instrumentation is not necessarily about changes in the individual, but rather changes in the measurement device or how the measure is used. One can make both threats implausible by ensuring that a control group or condition is included in the study that would show any testing or instrument effects without receiving the experimental manipulation or intervention. Thus, any instrumentation effect could be separated from the impact of the manipulation.

2.5: Additional Threats to Internal Validity

2.5 Summarize each of the additional threats to internal validity

Statistical regression, selection biases, attrition, and diffusion of treatments round out the major threats to internal validity.

2.5.1: Statistical Regression

Statistical regression refers to the tendency for extreme scores on any measure to revert (or regress) toward the mean (average score) of a distribution when the measurement device is re-administered.

If individuals are selected for an investigation because they are extreme on a given measure, one can predict on statistical grounds that at a second testing the scores will tend to revert toward the mean. That is, the scores will tend to be less extreme at the second testing. A less extreme score is, of course, one closer to the mean. That means individuals with very low scores and individuals with very high scores as a rule (but not necessarily everyone) will move to less extreme scores. This phenomenon is a threat to validity only if the changes in the group(s) in the study could be explained by simple regression rather than by some other interpretation such as the effect of the experimental manipulation.

In many areas of work (e.g., educational, clinical work, prevention), subjects are selected because they have high (e.g., for suicide risk, depression, antisocial behavior) or low scores (e.g., poor body image, self-esteem). One can expect that scores will change in the direction toward the mean of the overall sample whether that overall sample is included or not. This is not always the case, as we discuss later, but is of a concern, especially in intervention research where subjects are selected and screened precisely because they have an extreme score and warrant special attention. Their scores are likely to improve from one occasion to the next even if no intervention is provided!

It is worth elaborating statistical regression because it has other important methodological gems hidden in it. Consider for a moment we are going to assess many individuals for a study on anxiety. We administer a measure to many people and get their scores. Each participant's score can be considered for present purposes to have two parts:

1. Their true level of anxiety.
2. Error associated with unreliability of the measure, daily fluctuations in each person's normal variation in behavior, and no doubt other factors.

The error in measurement is reflected in the fact that scores from one testing to another are imperfectly correlated. Scores that are extremely high on one day are likely to be slightly lower on the next day or assessment occasion. Conversely, scores that are extremely low one day are likely to be slightly higher on the next testing.

Why do you think that happens?

The reason is that for many of the extreme scores the amount of error in the measure happened to be high (or low) and that error on the next occasion is not likely to be as high or as low. As a general rule, the more extreme the score, the more likely it is to revert in the direction of the group mean on subsequent assessment. Not every high score will become lower and not every low score will become higher, but on the average the scores at each of these extremes will revert or regress toward the mean.

2.5.2: Three Ways to Help Protect against Statistical Regression

Regression as a threat to validity can go up if extreme groups are selected. In clinical research they often are—we might want individuals who are high on anxiety (or some other domain of functioning). As we take the extreme group, we know that as a function of statistical regression they are likely to be less extreme when assessed again. If we are evaluating some intervention, we want to be sure that improvements that might have resulted from regression are not confused with improvements that resulted from treatment. That is, statistical regression is a threat to internal validity if it cannot be separated from the improvements due to treatment.

There are three ways to help protect against statistical regression as a problem:

1. Assign participants randomly to an experimental and control (or other condition). That way, regression if present will affect all groups, and one can see if the experimental manipulation or intervention led to changes beyond what was evident in the control group.
2. Use measures that are known to have high reliability and validity. The reason is that regression is a function of error of the measure. The greater the error, the more likely there will be regression. Stated another way, the correlation from one occasion to the next (test–retest reliability) is a good measure of error. Some measures (measuring height on one occasion and then 1 week later) have really high test–retest correlation. There is little error in the score, and regression is unlikely or minute. You do not get shorter or taller from one occasion to the next. (There is a little error in how one is standing.) Yet, the

measurement of height from one occasion to the next (e.g., a week apart) is highly correlated and almost exactly the same. That means three is little error and hence extreme scores are "real" and not filled with measurement error. In contrast, you use a psychological self-report measure that you made up or that is not well established. There is likely to be a lower test–retest correlation and then more error and more regression of extreme scores.

3. An excellent but rarely used strategy is to test everyone twice for the pretest and select only those individuals who were extreme on both occasions. Regression when it occurs is from the first to second assessment. Two assessments can be done before the study, and those whose errors contributed to their extreme scores are not likely to show extreme scores on two occasions due to regression. So people who score really high or really low on both assessment occasions are likely to have the characteristic of interest (e.g., anxiety) at their respective levels. This strategy is not used because it is not feasible. I mention it here because it helps to understand regression and how error of measurement is the culprit.

2.5.3: Selection Biases

A selection bias refers to systematic differences between groups before any experimental manipulation or intervention is presented.

Based on selection or assignment of subjects to groups, they already are different from each other in an important way that might contribute to or explain the results. At the end of the study, groups (e.g., experimental vs. control) may differ from each other but they may have differed even without the experimental manipulation. Obviously, the effects of an independent variable between groups can be inferred only if there is some assurance that groups do not systematically differ before the independent variable was applied.

In many ways, selection bias is the most obvious threat to internal validity and the one most frequently controlled in experiments. In an experiment, investigators routinely randomly assign subjects to conditions so that any subject characteristics (e.g., age, sex, and diagnosis) that may introduce a selection bias are dispersed among groups roughly equally or at least unsystematically. Of course, we would not have all women in one group and all men in another. Selection biases are not that stark as different sexes assigned to different groups, but if two or three times as many males were assigned to one group than another, there might well be a selection bias. Random assignment of subjects is the procedure commonly used to minimize the likelihood of selection biases, but as we discuss later that is no sure fire protection at all. Yet, random assignment of subjects to groups does make group differences not very plausible especially as the size of the sample increases. And, the goal in addressing a threat, whether selection bias or another one we have already discussed, is to make that threat implausible as an explanation of the results.

Selection bias often arises in clinical, counseling, and educational research where intact groups are selected, such as patients from separate clinics or hospitals and students from different classes and different schools. In prevention programs, for example, comparisons often are made among classes, schools, or school districts that receive or do not receive the intervention. Random assignment of classes or schools may not be possible for practical reasons (e.g., proximity of the schools in relation to the investigator, willingness of the school to have the intervention program).

Also, of course, random assignment of the children to different classes or schools is not an option as a general rule. Classes pre-formed before the researcher arrives and cannot be rearranged for research purposes. And, one cannot assume that groups seemingly equal really are. For example, we might have two third-grade classes, two fourth-grade classes, and two fifth-grade classes, and one class at each grade level is assigned to prevention program versus no program conditions. Yet, the classes may be at different schools, but wherever they are they were not composed by randomly assigning students to them. Thus, the project begins with special responsibility of the investigator to make implausible that selection (pre-intervention differences) might account for any group differences. There are options for making selection implausible. For example, there are special ways of matching subjects statistically on several variables and other ways of evaluating whether a particular variable that may influence the outcome in fact does by doing analyses within groups. More will be said later about addressing selection when we discuss designs in which random assignment cannot be used.

The discussion of selection bias to this point focused on experiments in which the investigator controls assignment of individuals or groups (e.g., classrooms, schools) to conditions and experimentally manipulates those conditions. Yet, not all studies are experiments. Often we are interested in understanding events, conditions, and processes where they occur in nature and intact groups are compared. For example, we want to know what a special group is like (e.g., depressed adolescents or soldiers with posttraumatic stress disorder). We begin with a select group. Here this is not selection bias as a threat to internal validity. The purpose of the study is to identify different groups and elaborate their unique characteristics. We will return to this because of other problems that can emerge (in relation to construct validity).

2.5.4: Attrition

Attrition or loss of subjects may serve as a threat to internal validity.

Loss of subjects occurs when an investigation spans more than one session. In clinical research, sometimes the study is conducted over the course of days, weeks, months, or even years. Intervention studies and longitudinal investigations that track individuals over time are primary examples. Some of the subjects may leave the investigation after the initial session or assessment, refuse to participate further, move to another city, or die. Yet, even if the study is just two sessions, attrition can be a problem. Some subjects who complete the first session may not come back for a second section. The problem is that the investigator went to the trouble of randomly assigning subjects to conditions and now some people are dropping out. The groups are no longer randomly comprised, and subjects, on the basis of variables we do not really know, elected to pull themselves out of the groups. We could easily have difficulty in detecting a selection bias now that the groups are no longer randomly comprised. We are now back to the possibility as a potential threat to internal validity.

Consider the following in a hypothetical two-group study in which participants serve for two sessions one week apart.

In group 1, let us save 85% of the subjects returned for the second session and in group 2 about the same percentage also returned. We might experience false security for a brief moment by saying even though there were dropouts, they were about the same number for each group. Actually the similarity of the number is not a critical feature here, and in relation to the threat to internal validity does not matter. The people who dropped out of each of the groups may not be random or be "identical" or equal but may have dropped out because of something special in each of the groups. So for example in a study comparing meditation and medication for anxiety, dropouts in the meditation group may not be like dropouts in the medication group. These are very different conditions, and those who drop out of each are not necessarily or even likely to be identical in all ways that might affect the results. The self-selection of attrition, patients left of their own accord, alters the composition of the group achieved through random assignment.

While attrition can be a problem in any study in which subjects need to return for additional sessions or complete measures over time, as one might expect the loss of subjects (how many drop out) is a direct function of time. Most of the subjects who will drop out tend to leave early, i.e., after one or two sessions. As the study continues (weeks, years), there continues to be a trickle of more people dropping out but at a slower rate than early in the study. We will discuss longitudinal designs much later in the text. In such longitudinal designs lasting years and sometimes decades, investigators often go to great lengths to keep in contact with subjects over the course of the study (e.g., contacting them at holidays, birthdays, periodic phone calls) and provide incentives (e.g., usually money) for completing assessments over the course of the project—all in an effort to avoid attrition and the impact attrition can have on conclusions of the study.

Attrition could be a threat to validity if there is any loss of subjects. Changes in overall group performance on the measures may be due to the loss of those subjects who scored in a particular direction, rather than to the impact of an experimental manipulation or intervention.

That is, the mean of the dropouts may be different from the mean of the rest of the sample so that changes in the mean may result from the loss of a select group of subjects. The plausibility of attrition as a threat can depend on how many were lost. If there were 100 subjects and 2 dropped out, clearly that is not as much of a problem as if 22 dropped out. We will return to the topic of attrition again because it is a threat to other types of experimental validity (external, construct, data-evaluation) that we have yet to discuss. There are strategies to deal with attrition too, and we will discuss those as well.

2.5.5: Diffusion or Imitation of Treatment

This threat can occur in any experiment where groups are exposed to different procedures, but is more likely to be a problem in intervention research (e.g., treatment, prevention, education). In these studies, it is possible that the intervention given to one group may be provided accidentally to all or some subjects in a control group as well. Obviously, one does not give the treatment to a control condition. And certainly if a subject is assigned to treatment, he or she ought to receive the treatment rather than the control condition.

Administration of treatment to the control group is likely to be inadvertent or accidental and, of course, opposite from what the investigator has planned. Yet, when this occurs, the effect will be to attenuate (dilute) the effects of treatment (since both groups received some treatment) and alter what the investigator concludes about the efficacy of treatment. Rather than comparing treatment and no-treatment conditions or two or more distinct treatments, the investigator actually is comparing conditions that are more similar than intended—they are blurred in what they received and that blurring is the diffusion. As a threat to internal validity, the effect of a diffusion of treatment is to equalize performance of treatment and control groups and thus reduce or distort effects of the intervention obtained

in the study. No differences statistically between groups at the end of the study or small differences may have as a plausible rival hypothesis that the treatment conditions (e.g., treatment vs. control, two or more treatments) spilled over into each other, or "diffused" and led to the pattern of the results.

Diffusion of treatment is not a trivial or infrequent problem and affects a range of areas. For example, years ago a special program was designed to decrease heart attacks among men (N = 13,000, ages 35–57) at risk for coronary disease (Multiple Risk Factor Intervention Trial Research Group, 1982). The intervention included personal dietary advice, drugs to control hypertension, advice to stop smoking, and exercise. Random assignment permitted comparison of this group with a control group that received testing (physical exams) but no special intervention (routine care). A follow-up 6 years after the program showed that the intervention reduced risk factors for heart disease but death rates due to heart disease were not statistically different between intervention and control groups. The absence of group differences has been interpreted to reflect a diffusion of treatment because subjects in the control group adopted many health-promoting practices on their own and also decreased their risk factors. (Actually, history, maturation, and diffusion of treatment could explain the absence of differences, and it is not easy to make the distinction here.)

Another facet of diffusion, no less significant, is that some cases in the intervention condition do not receive that intervention. In this situation, a diffusion of the no-treatment condition occurs, i.e., no treatment "spreads" to cases (e.g., individuals in a therapy study, classes in a prevention study) assigned to receive the intervention. It is possible that subjects did not show up for the program, that some oversight occurred, or that subjects were out ill and missed the program, or that the experimenters providing the intervention made an error and thought subjects were in the control or some other group. The net effect is the same, namely, where there is a diffusion of the conditions, the conclusions at the end of treatment are likely to be misleading. In this case, no treatment spilled over into the group that was supposed to receive the intervention.

> As a general rule, it is important to ensure that individuals assigned to a particular condition received that condition and only that condition.

The test of the experimental manipulation depends on ensuring participants received only the conditions to which they were assigned. In laboratory studies conducted in one session, often simple scripts of experiments or automated lab materials control the delivery of conditions. Even so, it is advisable to check to be sure that the conditions do not "diffuse" in any way.

2.5.6: Special Treatment or Reactions of Controls

This threat refers to a special circumstance in which an intervention program is evaluated and provided to an experimental group, but the no-intervention control group receives some special attention that can contribute to the results. That special attention poses a threat to internal validity if it is a plausible explanation of the findings. This is likely to occur in applied settings such as schools, clinics, and industry rather than in laboratory studies with college students. More explanation is required because this is more intricate than other threats we have covered so far.

Here is the usual scenario.

When a program is first proposed (e.g., to various schools or clinics), potential participants might be enthusiastic to participate. They learn that a special program (e.g., training of teachers or therapists; other services for students or patients) will be provided to address a problem of interest. For example, the program might be a special intervention to reduce bullying in the schools. As the program is described, participants may learn that through random assignment some schools will receive the special intervention and others will not. So far everyone is fairly happy; they have a chance to get a free intervention that will help with a significant problem they care about. Then random assignment is completed, and some schools are informed that they will not be receiving the intervention. That is, through the bad luck of the draw (random assignment) they do not get the program. From the investigator's perspective, the schools that do not receive the intervention are critically important. They serve as a control condition, and their assessment (e.g., before and after the period in which the program is implemented in the other schools) is essential to help make implausible such internal validity threats as history, maturating, testing, and statistical regression maybe other threats depending on how the schools were selected.

> How to keep the control schools in the study to avoid attrition and possibly the resulting selection biases? The control schools (staff, teachers) are now demoralized and may not want to participate now that they know they will not receive the intervention. After all, why complete all of the assessments if you are not getting any advantages of the intervention? Investigators may work hard to keep all the schools in the project to avoid selection bias from attrition as one source of a problem.

2.5.7: Additional Information on Reactions of Controls

Although the participants in the control schools may not receive the specific intervention of interest, they may

receive other services such as more money for school supplies, more materials for the classroom, some workshops (unrelated to bullying), and other such accoutrements just to keep them engaged. These are usually intended to redress the apparent inequality and to compensate for not providing the intervention.

From the standpoint of internal validity, however, the no-intervention group may be receiving an "intervention" in its own right that obscures the effect of the program provided to the experimental group. That special treatment is different from "no treatment" and might interfere with interpretation of the outcome. The special treatment is not exactly diffusion of treatment because essential ingredients from one group do not spread (diffuse) to another, but it is like that. Special attention and hovering do spill over into both groups and that part is like diffusion. After all, the no-intervention group receives something special. At the end of the study, no differences between the groups might be due to comparing two interventions that worked rather than the ineffectiveness of the main program.

There is another influence that is part of special treatment as a threat to internal validity. When participants are aware that they are serving as a control group, they may react in ways that obscure the differences between treatment and no treatment. Control subjects may compete with the intervention subjects in some way on their own, i.e., without the investigators providing anything special. For example, teachers at control schools who learn they are not receiving the intervention (e.g., to improve student academic performance) may become especially motivated to do well and to show they can be just as effective as those who receive the special treatment program. On the other hand, rather than trying extra hard, controls may become demoralized because they are not receiving the special program. The controls may have experienced initial enthusiasm when the prospect of participating in the special intervention was announced, but their hopes may be dashed by the fate of random assignment. As a consequence, their performance deteriorates. By comparison, the performance of the intervention group looks better whether or not the intervention led to change. In short, in these scenarios, we have a group that is not quite "no treatment" to provide a baseline of change to evaluate history, maturation, and testing along. Rather, they receive some intervention or reacted in a way to be a self-imposed intervention.

> Awareness of participating in an experiment can influence both intervention and control groups. From the standpoint of internal validity, a problem arises when this awareness differentially affects groups so that the effects of the intervention are obscured. At the end of the study differences between treatment and control subjects, or the absence of such differences, may be due to the atypical responses of the control group rather than to the effects of the intervention. The atypical responses could exaggerate or attenuate the apparent effects of treatment.

In clinical treatment research (e.g., a study on the treatment of depression) work in clinical settings, some new intervention often is compared to "treatment as usual," that is the intervention that is provided in that setting. Among the reasons is that all cases receive something that is or seems legitimate. One does not have to worry about demoralization from no treatment and attrition caused by that. Using treatment as usual at least ensures that everyone (whether treatment or control) receives some viable intervention and this makes less plausible, but does not eliminate, the possibility that specialness of the intervention could explain the results. Similarly, in experimental laboratory studies, the group that receives the experimental manipulation usually can be compared with another group that receives something but that something may not be expected to have an effect. There is no special treatment, extra motivation, or demoralization under such circumstances.

2.6: When and How These Threats Emerge

2.6 Review the four main circumstances of potential threats to internal validity

Ideally, it would be instructive to select a single study that illustrated all of the threats to internal validity. Such a study would have failed to control for every possible threat and would not be able to attribute the results to the effects of the independent variable. A study committing so many methodological sins would not be very realistic and would not represent most research efforts in which flaws are committed only one or a few at a time. Thus, detailing such an ill-conceived, sloppy, and uncontrolled study would have little purpose. (It would, however, finally give me a forum to present the design and results of my dissertation.) The threats to internal validity can raise problems under four main circumstances, and these are useful to illustrate and be wary of.

2.6.1: Poorly Designed Study

> A poorly designed study is one in which from the outset we know that many threats will be plausible.

While this is not common in psychological research, one would be stunned to see how often this occurs. The most common and flagrant example is a single pre-post design. Let us be brief because understanding this one will not hone your skills in moving toward being black-belt methodologists.

In this case, a single group is selected and tested before some manipulation or intervention. Then the manipulation is given followed by posttest. Low and behold the posttest is significantly different (statistically) from the pretest—people got better! In this case, history, maturation, and testing all are quite plausible rival hypotheses. If the subjects were selected because they have a problem of some kind (e.g., extreme score), statistical regression might be added to the mix. Changes in a single group from one test occasion to another do not require saying it was the manipulation. Threats to validity are plausible. They are also parsimonious because they can explain similar effects in many studies and there is no need to add another concept (the manipulation) of this one study.

It is not difficult to find examples. As a brief illustration, in one study 59 hospitalized patients were exposed to pet therapy (Coakley & Mahoney, 2009). Patients were hospitalized for a variety of physical problems (e.g., cancer, asthma, AIDS, heart failure, diabetes, coronary artery disease, gastrointestinal bleeding, amputation, hysterectomy, and other conditions). The goal was to reduce stress-related outcomes by exposure to pet therapy, which included a visit with a dog or dog handler in the patient's room. From pre-to-post-treatment assessment, the patients showed statistically significant decreases in pain, respiratory rate, and negative mood state and a significant increase in perceived energy level. Sounds great. Yet, parsimony and plausible rival hypotheses force us to say history, maturation, and testing were not ruled out and can easily explain the results! More strongly stated, there is no evidence in this study that the pet contributed to improvement; it may have, but this study cannot tell us that.

> One might ask, why should we care whether pet therapy made a difference, as long as the patients got better? We care greatly because the elderly usually do not receive the services they need and interventions to address physical pain and mood are sorely needed. We want to know what interventions make a difference so that we can apply them widely.

We also want to know it is the intervention, so we can analyze that further, tweak it to make it more effective, perhaps make changes to improve the impact and durability of the change, use it for other populations than the elderly, and so on. The first step—a study that rules out threats to validity. We need to know whether it is the intervention that can account for the changes once threats to internal validity are controlled.

When one is beginning a research career or beginning a new area of study, it is often useful to do a pilot study of one group. A pilot study is like a dress rehearsal of the "real" study.

Here the goal is to get a feel for the experimental manipulation, how to do it, to ask clients about desirable and objectionable procedures, to evaluate outcome and preliminary results, and so on. In this case, a pilot study without a control condition is fine to work out details, feasibility, and so on. Pilot work is intended to help develop a controlled study from which inferences about the intervention effect can be drawn and pretest-posttest of one group without controls is fine.

2.6.2: Well-Designed Study but Sloppily Conducted

Let us say this is a well-designed study. Participants were assigned randomly to groups:

- One group gets something (e.g., treatment, induction of some emotional state)
- The other group gets nothing or something different (e.g., no treatment, induction of no special emotion)

So far so good, but sloppy procedures can easily raise the prospect of a threat to internal validity.

Diffusion of treatment (or conditions) is a likely candidate. This can occur if the investigator does not monitor the manipulation to ensure that it is delivered correctly to each group and there is no mix-up, spillover, or lapses. There may be no differences between conditions at the end of the study. One plausible interpretation might be that the conditions diffused into one another. One would like data (e.g., from tapes of sessions, from checklists completed by observers who watched perhaps a random sample of sessions, or records of the experimenters who ran the conditions). These data could reassure us that the conditions were properly administered.

> Within a study, it is critically important to ensure that the different conditions are implemented exactly as intended. In laboratory studies with college students receiving some audio or video instructions, this is fairly easy to do. Similarly, in computerized or online Web-based studies, video material presented to different groups can ensure standard administration of the conditions and correctly. But in studies on applied problems and in applied settings (e.g., clinics, schools, counseling centers) or with interventions carried out over time, this is not so easy.

One often cannot standardize the manipulation or intervention (e.g., with recordings or videos), and the persons who apply the intervention (e.g., therapists, teachers, counselors) have many tasks and responsibilities that may pull them from focusing rigidly and meticulously on delivering the condition faithfully. Even so, investigators can do more. In most studies of therapy (even the most rigorous studies), researchers do not measure how faithfully or correctly the interventions were administered (Perepletchikova, Treat, & Kazdin, 2007). This can make interpretation of the results very difficult. Were the treatments really no different in their

effects, or was the sloppiness of administration (diffusion, lapses in even delivering the treatment) the culprit? We will return to the matter of unmonitored manipulations or interventions because neglect here can undermine a study in other ways.

2.6.3: Well-Designed Study with Influences Hard to Control during the Study

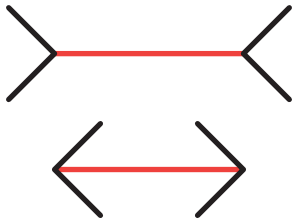
Threats can emerge because of circumstances that are not so easily controlled by the investigator. Here again, the study might be well designed as in the previous instance. Yet, if there are two or more meetings or sessions in the experiment, attrition or loss of subjects may occur and systematically bias the results. Attrition can be controlled a bit by the investigator over the course by keeping the study brief or providing incentives (e.g., money or a chance to win some electronic device) for completing the final session. In grant-funded research, often money is available to provide a sum (e.g., few hundred dollars) if clients complete the final intervention session and the post-intervention assessment battery. Alternatively, if subjects were in a control group, sometimes they are promised access to the intervention without cost but only after they complete the second "post" assessment for the period in which they did not receive the intervention.

Even with the best strategies, in treatment, prevention, and longitudinal studies, some subjects invariably will drop out, and if there are two or more groups, selection biases may enter into the study and threaten internal validity. There are statistical options to handle this (e.g., intent-to-treatment analyses, we will mention later), but there is only so much the investigator can do. But the point remains, a selection bias may emerge during a study and that is not always easily controlled.

2.6.4: Well-Designed Study but the Results Obscure Drawing Conclusions

Consider a final case in which internal validity threats can emerge. Let us begin again with well-designed study including random assignment of participants to two (or more) conditions (some manipulation or two or more treatments). At the end of the study, the results show that the conditions were equally effective. Figure 2.1 gives a hypothetical example. Key to the example is that there are two groups and both groups changed. (The same point would apply if there were three groups with this data pattern.) Let us say pre to post change was statistically significant for each group, which means that the amount of change each group made met the criterion for statistical significance. Yet, the groups were not different from each other at the end of treatment, which means that between-group differences did not meet statistical significance.

Figure 2.1: Muller-Lyer Illusion



Now the investigator may wish to say, "Both of the treatments or manipulations were effective. After all, both groups showed a statistically significant improvement from the beginning to the end of the study." Not so fast. This is not the conclusion permitted by the results. The design was fine at the outset, but the results make any conclusions ambiguous. History, maturation, testing, and maybe statistical regression (if subjects were screened in any way to select more extreme scores) rear their ugly heads and can completely explain the results. The proper conclusion by the investigator: "There is no evidence that my manipulations (interventions) made any difference whatsoever; the changes over time could be due to any of several threats to internal validity or to the intervention."

This is an optical illusion with multiple variations. The horizontal lines (or shaft) of the arrows are the same length. When the ends of the shaft point outward (top portion), the shaft often is perceived as longer than when the ends of the shaft point inward (bottom portion). The illusion is named after the person who is credited with its identification in 1889, a German sociologist named Franz Carl Müller-Lyer (1857–1916).

Whenever you see a graph that resembles what was evident in Figure 2.1 in which all groups show parallel lines (or approximations of that) from one assessment occasion to the next, say to yourself, history, maturation, and other threats could readily explain this. And, this can be said, if the groups are any combinations of various interventions and control conditions. This can be frustrating for any investigator because the study seemed so great to begin with, and an investigator is not responsible for the pattern of the results (but that is not always true based on the specific manipulations selected). All that said, threats to internal validity remain plausible and parsimonious.

> Does this ever happen? Yes and often. For example, in one study an intervention (sending postcards regularly) was

compared to treatment as usual in an effort to reduce self-poisonings of individuals once they left the hospital after inpatient treatment for that problem (Carter, Clover, Whyte, Dawson, & D'Este, 2013). At the end of the intervention period and over the course of the follow-up, there were no differences between groups on subsequent hospital admissions for self-poisoning.

Both groups showed a reduction in the number of incidents of self-poisoning over time.

Were both interventions effective? Maybe but we cannot tell because improvements over time could readily be explained by threats to internal validity.

As another example, three groups (cognitive therapy, attention placebo, treatment as usual) in a multi-city study focused on reducing risky sexual behavior and substance use in men with HIV/AIDS (Mansergh et al., 2010). Rates of the outcome measures decreased over time through posttreatment and follow-up period, but groups were no different. Threats to internal validity (e.g., maturation, testing) are plausible explanations of the results here as well. With both examples, we cannot say that a particular threat explained the changes. Yet, they are plausible explanations, and hence we may not need to use "treatment" as an explanation for the sake of parsimony. Said another way, we cannot conclude from these studies that treatment led to change.

2.7: Managing Threats to Internal Validity

2.7 Discuss the importance of determining the relevance of a threat to internal validity in order to manage it

The first task in designing an experiment is to go through the list of potential threats to evaluate if the threat is relevant at all and if it is relevant whether it could be a problem. If there is no selection of extreme groups or screening criteria used to select subjects, regression is not likely to interfere with the conclusions. If there is no repeated testing, then testing effects are not relevant and so on. Yet, it is useful to consider each threat explicitly to judge whether it could be a problem.

If threats are potentially relevant, often they can be readily controlled. Usually one does not need a specific strategy to deal with each threat. The threats come in packages and can be controlled that way as well. History, maturation, and testing, for example, are handled by using a control or comparison group that is randomly comprised. Random assignment and assessment of the control subjects without the experimental manipulation takes these threats off the table.

Randomly assigning subjects usually handles selection biases.

Some threats to internal validity emerge as the study is conducted. Among them is attrition where subjects drop out and raise the prospect of a selection bias and cancel out the random composition of groups. Of course most studies are one session, and attrition is not the issue there. As more sections are part of the experiment, the prospect increases that attrition will be a problem. Different strategies are used to retain individuals in the study. Often monetary incentives are provided for completing the final session and completing any incentives. In longitudinal studies, often conducted over years and decades, the research team often stays in close contact with the participants (e.g., phone calls, notices at birthdays and other holidays, newsletters) to keep them committed to the project. At the end of the study, some subjects may have dropped out.

Several statistical approaches to attrition have been developed and provide useful strategies to complement active efforts to minimize attrition. Statistical methods allow researchers to identify the likely bias that attrition introduces into the data and the conclusions that would be warranted if the lost subjects had improved, remained the same, or became worse.

Diffusion of treatment also can emerge as the study is conducted. The conditions (groups) may have procedures that blur (diffuse) into each other, so the experimental group did not really receive the condition as intended nor did the control group.

Some in the experimental group may not have received the manipulation and someone in the control group may have. If there are complex experimental procedures and procedures cannot be automated (e.g., prerecorded, presented on a laptop), it is worthwhile monitoring implementation of the conditions to ensure that each condition was executed correctly and that the conditions did not in any way blend into each other. Among the alternatives is that sessions might be taped in some way and then they can be checked to evaluate the procedures. A checklist might be made to evaluate the tapes to ensure that critical features of the manipulation or session took place and features of the other condition did not. Diffusion of treatment has been of special concern in intervention programs where two or more interventions are compared. In such programs, one wants to be assured that the interventions were administered correctly. The term "integrity of treatment" or "treatment fidelity" is used in the context of intervention programs (Perepletchikova et al., 2007). For now, it is important to evaluate if there is any possibility in a study

that the conditions will be more similar than intended or aspects of one condition will accidentally blur into the other condition. This is a huge threat to validity because the conditions will look quite similar in the effects they produce not because their effects really are similar but because the conditions overlapped in ways that were not intended.

2.7.1: General Comments

Threats to internal validity are the major category of competing explanations of the conclusion that the intervention (manipulation or experimental condition) was responsible for group differences or changes. If a threat to internal validity is not ruled out or made implausible, it becomes a rival explanation of the results. Whether the intervention or particular threat to validity operated to cause group differences cannot be decided, and the conclusion about the intervention becomes tentative. The tentativeness is a function of how plausible the threat is in accounting for the results given the specific area of research.

Some threats may be dismissed as implausible based upon findings from other research that a particular factor does not influence the results. For example, consider as a hypothetical example, use of an experimental medication for 20 patients in the last stages of a terminal disease. Here there is no control group, which often is the case in developing an intervention in the early stages. The results may show that some or most patients did not die in say the 2-month period that normally would occur at this point in the disease. Indeed, patients may have lived for a mean of 8 months. We still need controlled trials here, but the example is good enough to illustrate the point.

> *History, maturation, and testing (being alive is the measure), and statistical regression (being selected because they were extreme on the dimension of interest) probably are not very plausible as threats to internal validity.*

The reason here is that decades of research and clinical care have shown that death is the usual course. The medication may be more plausible as an explanation for extending survival.

From a practical standpoint, it is important and useful for an investigator to decide in advance whether the investigation, when completed, would be open to criticism to any of the threats and, if so, what could be done to rectify the situation. In the examples cited where history, maturation, and regression were plausible, a no-treatment condition or an intervention as usual (what is routinely done) to rule out each of these threats might have been included. As I noted, not all threats to internal validity can be considered and completely averted in advance of the study. Problems may arise during the study that later turn out to be threats (e.g., instrumentation or attrition), and others might stem from the pattern of results (all groups improve at the same rate). Even so, with many problems in mind prior to the study, specific precautions can be taken to optimize the clarity of the results.

2.8: External Validity

2.8 Define external validity

> External validity refers to the extent to which the results of an investigation can be generalized beyond the conditions of the experiment to other populations, settings, and circumstances.

External validity encompasses all of the dimensions of generality of interest. Characteristics of the experiment that may limit the generality of the results are referred to as threats to external validity.

2.9: Threats to External Validity

2.9 Summarize different threats to external validity

Threats to external validity constitute questions that can be raised about the limits of the findings. It is useful to conceive of external validity as questions about the boundary conditions of a finding. Assume that a study has addressed the issues of internal validity and establishes a relation between an experimental manipulation (e.g., way to cope with pain in some laboratory paradigm using introductory psychology students) and outcome. One is then likely to ask:

- Does this apply to pain that is normally encountered in every life that is not part of some laboratory setup?
- Does it apply to other groups of persons (e.g., diverse ethnic or racial groups, young children, the elderly, and so on), to other settings (e.g., clinics, day-care centers), or to other geographical areas (e.g., rural, other countries)?
- What are the boundaries or limits of the demonstrated relationship?

Stated another way, one can discuss external validity in terms of statistical interactions. The demonstrated relation between the independent and dependent variables may apply to some people but not others or to some situations but not others, i.e., the independent variable is said to interact with (or operates as a function of) these other conditions. For example, if the finding is obtained with women but not men, we say that the intervention interacts with subject gender. Also, one could say the effects of the treatment are moderated by

gender. We will return to the concepts (interactions, moderators) later apart external validity. The factors that may limit the generality of an experiment usually are not known until subsequent research expands upon the conditions under which the relationship was originally examined. The manner in which experimental instructions are given; the age, ethnicity, sex, or gender identity of the subjects; whether experimenters are from the general population or college students; the setting in which the experiment is conducted; and other factors may contribute to whether a given relationship is obtained.

> One has to be careful in using threats to external validity as a means of challenging findings of a study. These threats often are used as superficial criticisms of an investigation.

That is, one can always criticize a study by saying, the finding may be true, but the investigator did not study subjects who (and now you "fill in" some subject attribute) were much older or younger, were of this or that ethnicity, were from another country, or who were bald, bipolar, brainy, brash, or had some other characteristic beginning with "b." These tend to be superficial criticisms when they are posed cavalierly without stating precisely why one would expect findings to be different as a function of the characteristic selected. In fact, the generality of experimental findings may be limited by virtually any characteristic of the experiment and the subject characteristics noted previously might well be plausible. There is some responsibility of the individual who poses the concern to explain in a cogent way why it is a threat.

A threat to validity must be a plausible factor that restricts generality of the results.

2.9.1: Summary of Major Threats

Also, parsimony guides us and in this context directs us to the view that we do not need to introduce additional concepts (subject characteristics of all sorts) to explain a finding without evidence that the additional information is needed. This does not mean that the finding applies to everyone, for all time, and in all settings. It just means we do not go wild introducing complexities until required (based on clear theory and, even better, on clear data) to do so. Although we cannot know all of the boundaries of a finding (i.e., their limits), several characteristics, or threats to external validity, can be identified in advance of a particular study that might limit extension of the findings, but again to be threats they must be plausible. That is, we have to have a little theory or prior findings to suggest the threat is quite possible. More concretely, we do not begin with, yes but the findings might not apply to this or that cultural group. The onus is on us if we state this, to continue the sentence "because . . ." and then to note in a cogent finding specifically why one would expect differences. Table 2.3 summarizes major threats to external validity to provide an easy reference, and let us discuss them as well.

Table 2.3: Major Threats to External Validity

Specific Threat	What It Includes
Sample Characteristics	The extent to which the results can be extended to subjects or clients whose characteristics may differ from those included in the investigation.
Narrow Stimulus Sampling	The extent to which the results might be restricted to a restricted range of sampling of materials (stimuli) or other features (experimenters) used in the experiment.
Reactivity of Experimental Arrangements	The possibility that subjects may be influenced by their awareness that they are participating in an investigation or in a special program. The experimental manipulation effects may not extend to situations in which individuals are unaware of the arrangement.
Reactivity of Assessment	The extent to which subjects are aware that their behavior is being assessed and that this awareness may influence how they respond. Persons who are aware of assessment may respond differently from how they would if they were unaware of the assessment.
Test Sensitization	Measurement in the experiment may sensitize subjects to the experimental manipulation so that they are more or less responsive than they would have been had there been no initial assessment. This prospect is more likely if there is a pretest and that pretest is one that alerts awareness that assessment is going on and what the focus of that assessment is.
Multiple-Treatment Interference	When the same subjects are exposed to more than one treatment, the conclusions reached about a particular treatment may be restricted. Specifically, the results may only apply to other persons who experience both of the treatments in the same way or in the same order.
Novelty Effects	The possibility that the effects of an experimental manipulation or intervention depend upon the innovativeness or novelty in the situation. The results attributed to the experimental manipulation may be restricted to the context in which that is novel or new in some way.
Generality Across Measures, Setting, and Time	The extent to which the results extend to other measures, settings, or assessment occasions than those included in the study. There is a reason to expect that the relations on one set of measures will not carry over to others, or that the findings obtained in a particular setting would not transfer to other settings, or that the relations are restricted to a particular point in time or to a particular cohort—these would be threats to external validity.

2.9.2: Sample Characteristics

The results of an investigation are obtained with a particular sample. A central question is the extent to which the results be generalized to others who vary in age, race, ethnic background, education, or any other characteristic. In research, there are different types or levels of concern in generalizing from one sample to another.

> **Human and Nonhuman Animals:** One concern is the extent to which findings from nonhuman animal research can be extended to humans. For example, this concern has emerged in laboratory animal research where experimental manipulation of diet (e.g., consumption of soft drinks or a particular food in rats) is shown to cause cancer.

Assume that the findings are unequivocal. We, as humans, want to know whether the results generalize from this sample to us. Sample differences are quite plausible because of the multiple cancer-related factors that may vary between laboratory rats and humans. The laboratory rats are given heavy diet of a soft drink, and humans normally consume some significantly lower proportion of their diet as soft drinks. The results may not generalize to subjects (humans) whose diets, activities, metabolism, longevity, and other factors differ. Also, special features of the subjects (rats of a particular species) may have made them differentially responsive to the intervention and hence restrict generality across species.

Generality is not an inherent problem in the leap from nonhuman animal to human research. Just the opposite, many of the major advances in psychology (e.g., learning), biology (e.g., brain functioning, genetic transmission, understanding HIV), and medicine (e.g., vaccination effects, surgeries) have derived from the fact that there *is* considerable generality across species. Also, in much of basic research, the specific nonhuman animal that is selected for study is based on some similarity with the system (e.g., immune, blood, digestive) of nonhumans and humans or the ability to study processes central to a human condition. For example, schizophrenia is a serious mental disease that has pervasive impact on psychological functioning (e.g., reward learning, memory, perceptual discrimination, object-place learning, attention, impulsivity, compulsivity, extinction, and other constructs).

Measures of these impairments in these functions have been developed in rodents to study processes likely to be applicable to schizophrenia (Bussey et al., 2012). We will discuss animal models later, but research often looks at identifying critical mechanisms that might be involved in a problem where applicability of key findings to humans is already known.

For example, exposure to low levels of lead among children is associated with hyperactivity, deficits in neuropsychological functioning (e.g., verbal, spatial ability), distractibility, lower IQ, and overall reduced school functioning in children. These effects continue to be evident several years later (see CDC, 2012a). Ingestion of leaded paint (e.g., children eating paint chips), fumes from leaded gasoline, and residue of leaded pipes from which water is consumed have been key sources of exposure, before changes in each of these sources have been made (e.g., shift to unleaded gasoline).

Lead collects in the bones and is concentrated in the blood and in high doses can cause seizures, brain damage, and death.

Several studies with humans have established the deleterious and enduring effects of low levels of lead exposure among children. Apart from several studies of humans, animal research has elaborated critical processes involved including the effects of low lead levels in rodents and monkeys on brain activity and structure (e.g., neurotransmitter activity, complexity in dendrite formation, inhibition of the formation of synapses, cognitive tasks, and how lead operates) and hence elaborates how learning and performance are inhibited (see CDC, 2012a). Capitalizing on this research, many countries have lowered lead exposure and the benefits have been reflected on increases in child IQ (Jakubowski, 2011). There is likely to be great generality of these findings in terms of the brain structures and functions affected across cultures, ages, and so on.

The lead example conveys areas where nonhuman animal research has been pivotal in elaborating findings immediately pertinent to humans. It would be a disservice to animal research and the issue of generality of findings to leave the matter here. The value of animal research among several scientific disciplines including psychology does not stem from immediate generality. Much of the understanding of basic processes (e.g., brain functioning) stems from animal research. Also, animal research often conveys what can happen in principle and raises questions about mechanisms of action and more will be said about that later.

2.9.3: College Students as Subjects

> **Heavy Reliance on Undergraduates as Subjects:** Another concern about sampling and external validity relates to the frequent use of college students as subjects.

Much of psychological research includes laboratory studies of critical topics (e.g., aggression, depression, memory) where students are brought into the laboratory for a brief experiment. How could the results be generalizable? College students as subjects have been referred to as WEIRD, an acronym for Western, Educated, Industrialized, Rich, and from Democratic Cultures (Henrich, Heine, & Norenzayan, 2010a, b). Evaluation of research suggests that approximately 67% of psychology studies

in the United States rely on undergraduates as subjects (Arnett, 2008).

> There are reasons to believe and supportive data to indicate that WEIRDos, as they are called, do not necessarily represent individuals from other cultures in fundamental ways in such areas as attributions, reasoning style, personality, perception, and others.

For example, recall from undergraduate psychology the famous Muller-Lyer Illusion (see Figure 2.1). Examination of scores of cultures indicates that not all people see the line in the top part of the figure as longer than the line in the latter. Although the lines are the same length, individuals from other cultures see the difference as much smaller than do WEIRDos. The broader lesson is that we often assume that we are investigating fundamental processes even in such core areas as perception and that these will have broad generality. In fact, that has been readily challenged now that we know more about strong cultural influences that can greatly influence generality as a result.

There are two levels to consider this point:

1. For psychology as a field, the heavy reliance on college students really does jeopardize generality of many of our findings. We do not know the extent of the problem because studies with college students are rarely replicated with other populations. Yet, we know that there are important cultural differences in how people view the world and see their place in nature (e.g., in relation to other and other beings) and that such frameworks affect memory, reasoning, perception, and perspective taking (Atran, Medin, & Ross, 2005; Bang, Medin, & Atran, 2007; Wu & Keysar, 2007). Moreover, cultural differences and preferences are mediated by variation in neural substrates, as demonstrated by brain imaging (functional magnetic resonance imaging [fMRI]) studies (e.g., Hedden, Ketay, Aron, Markus, & Gabrieli, 2008).
2. For our own individual studies, we only need to consider whether generality across samples is a priority for the study. In many studies, this is not necessarily a concern. First, we want to show whether there is a relation between variables and generalizing at this initial point may not be critical. In the case, sample selection is still important. We want to select the sample that is likely to show the effect, at least in this initial study. Generality of the finding at this point may not be something of concern. Even so, in light of recent research, it is useful to keep in mind that college students, the "standard" or most frequently used sample for research, may have special features that limit generality.

> In recent years, the heavily reliance on college students has been complemented by Internet studies that rely on nonstudent populations.

Researchers increasingly recruit subjects from Web sites such as Amazon Mechanical Turk (www.mturk.com/mturk/welcome) or Qualtrics (http://qualtrics.com/), which yields a broader range of individuals in terms of age, education, and other demographic characteristics than college students (e.g., Buhrmester, Kwang, & Gosling, 2011). Comparison of many findings indicates that college student and Internet samples show quite similar effects and the worry of generality of findings from one group to the other may be exaggerated. Such extensions are all to the good in terms of external validity. Also, as our own culture has become richer in minority group representation, the cultural differences can be studied further to evaluate the factors that influence generality of a finding.

2.9.4: Samples of Convenience

> **Samples of Convenience:** Related to the use of college students, occasionally there is concern about the using of other samples of convenience. This refers to the selection and use of subjects merely because they are available.

Obviously, a sample of subjects must be available. However, occasionally subjects are selected because they are present in a convenient situation (e.g., waiting room, hospital ward) or are available for a quite different purpose (e.g., participation in another experiment that requires a special population). An investigator may use an available sample to test a particular idea or to evaluate a measure he or she has just developed, but the sample may not be appropriate or clearly the one best suited to the test.

The most common use of samples of convenience is in situations in which a sample is recruited for and well suited to one purpose. As that study is begun, the original investigators or other investigators realize that the data set can be used to test other hypotheses, even though the original sample may not be the sample that would have been identified originally if these other, new purposes were the central part of the study. When samples of convenience are used, the onus is on the investigator to evaluate whether unique features of the sample may contribute to the results. The use of a highly specialized population that is selected merely because it is convenient raises concern. The specialized population and the factors that make them particularly convenient may have implications for generalizing the results.

As an extreme case, the sample may be recruited because of meeting criteria related to clinical dysfunction (e.g., use of illicit drugs, excessive consumption of alcohol) in keeping with the goals of a study. Another study is added on to that by perhaps adding a measure or two to evaluate depression, personality style, or some other domain not part of the original study. Utilization of the sample in novel ways is fine and often creative. However,

it may be appropriate at the end for the reader or reviewer of the report to ask, "why this sample?" How will the results pertain to the sample one cares about (e.g., people not recruited for some very special purpose)? Samples of convenience appropriately raise these concerns.

2.9.5: Underrepresented Groups

> **Underrepresented Groups:** A broad concern about the generality of findings from one sample to another pertains to the limited inclusion of women and underrepresented and minority groups as research participants. Historically, women and various ethnic groups were not extensively studied in the context of many topics in the biological, behavioral, and social sciences in the United States (and elsewhere).

The point about ethnicity was made dramatically in a book entitled, *Even the Rat Was White* (see Guthrie, 2003). Many studies (approximately 40%) did not even report ethnicity of the sample, so the scope of the problem is not easily assessed when the issue codified several years ago (Case & Smith, 2000).

Major efforts have been made to controvert this serious underrepresentation of minority groups, both of not including and of not reporting characteristics, of the sample. For example, federal funding agencies (e.g., National Institute of Mental Health) require including and specifying clearly what groups will be included in a study, and if a particular group is to be neglected, a firm rationale is needed. Even so, it is clear that many groups have not been routinely included in research. In relation to proportion of the population in the census (in the United States), some groups have been overrepresented in research (e.g., African Americans) and others have been quite underrepresented (e.g., Hispanic Americans, Native Americans). A critical issue is the extent to which findings may be restricted to those groups included in research.

I have mentioned cross-cultural differences previously in the comments about undergraduates as subjects and how findings may not represent what is obtained from other types of subjects (non-WIERDos). In the context of treatment of psychiatric disorders, we also know the important and ethnic differences and possible mechanisms involved. For example, responsiveness to medication for psychiatric disorders varies as a function of ethnicity (Lin & Poland, 2000). There are important reasons to expect ethnic differences in part because diverse groups differ in concentrations of various enzymes that influence metabolization of drugs. Many of the enzyme concentrations seem to be genetically controlled and perhaps emerged in defense to toxins (e.g., exposure to plants, pollen, and infection) in the environments that differ as a function of the respective geographical areas of origin for the different ethnic groups. Absorption and rate of metabolizing drugs that in some way utilize these enzymes can vary significantly and hence serve as the basis for expecting ethnic differences in response to medication (e.g., among African Americans, Asians, Caucasians, Hispanics, and Saudi Arabian adults). From a clinical perspective, this means that a recommended dose (e.g., of some antidepressant and antianxiety medications) for members of one group can be an overdose or underdose for members of another group. From a methodological perspective, this means that findings from a study with one ethnic group might not generalize to another group whose enzyme profile in relation to processes involved in a particular medication is known to differ from the profile of the sample included in the study.

The focus on underrepresented and ethnic minority groups in the United States raises broader research issues. As psychologists and social scientists, we are concerned with the generality of our findings across groups, especially those groups that have been neglected and underrepresented. Yet, the scientific agenda is much broader. We are interested in people of the world, many cultures, and many subgroups within a culture. Within our own culture, including women and men is strongly advocated in research proposals. Yet, not such systematic attention has been given to sexual identity, and there is no reason to not give that similar attention. The basis for saying that has to do with the prospect that sexual identity might well moderate (influence) many findings and studies of individuals with one particular type of identity may not generalize to others. Of course, we need to pose why that might be true in any given instance, but the topic has been fairly neglected.

> We wish to know the extent to which findings extend to these diverse groups, and principles or processes that can explain how culture influences affect, cognition, behavior, development, the family, and so on.

We would like to know about processes that explain generality of findings or lack of generality of findings among diverse groups and cultures in part because every key or major finding could not be studied with all different ethnic groups and indigenous populations in the world. This is not a reason to restrict research to one or a limited number of groups. Quite the opposite. However, the comments also convey that extending research to different groups is not an end in itself. Rather, the goal is to understand the processes, sources of influence, and factors that might dictate why or how a finding is one way in this context but another way in a different context. As these comments suggest, external validity is not just a topic of methodology, but raises important substantive questions about research findings, the factors that may influence them, and fundamental processes or mechanisms on which group differences are based.

When the study of the generality of findings across groups or samples serves as the basis for research, it is important to go beyond the basic question of whether or the extent to which prior results also apply to a new group. As a basis for new research, it is very useful to identify theoretical issues or to propose mechanisms or processes that would lead one to expect differences in the findings across previously studied and to-be-studied samples. Understanding processes or factors that may mediate (account for) differences of various samples is especially valuable not only for theoretical reasons but for practical ones as well. In the approximately 200 countries of the world, there are thousands of ethnic cultural groups, within individual countries often many groups (e.g., Chad in Africa, noting 200 different groups; www.infoplease.com/ipa/A0855617.html). In short, there are more groups than can ever be studied to test the generality of all or all major findings, whatever these findings would be. In the United States, the Census (2010) recognizes a limited number of groups:

- White, Black, or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Other Pacific Islander

2.9.6: Additional Information on Underrepresented Groups

However, a separate question is asked about Hispanic or non-Hispanic so that the 5 categories x 2 options yield groupings. Yet, these groupings greatly underestimate and misrepresent ethnicity and culture.[4] Moreover, in the United States, the most recent census places multiracial individuals as the fastest growing ethnic group. And of course studying this group raises challenges because multiracial group masks enormous richness and heterogeneity rather than "one group" that ought to be studied.

There are excellent opportunities to study generality of findings. Challenging the generality of a given finding or testing the generality in a subsequent study ideally has a strong conceptual or empirical basis for believing the original finding will not generalize.

Parsimony is where we begin, namely, that the original finding stands in explaining how variables are related across all groups.

It is then incumbent that someone who challenges that shows why that simple view cannot account for additional demonstrations.

As often as not, it may be important to show that findings do generalize. When one is testing the generality of findings to a new type of sample, it is important to make the case that there is an interesting reason to suspect differences among groups. A weak rationale for any new study is that, "whether the finding generalizes to this population has not been studied before." This is weak unless strongly bolstered by theory, data, or brilliant reasoning that suggests why lack of generality would be expected. This type of study is a much more significant contribution than merely assessing whether effects of prior research generalize to a new set of subjects who vary in some way from the original sample. There is a compromise position from merely replicating a finding to jumping from one sample to the next. That is to do the study including two (or more samples) and make predictions of differential responsiveness and why. That study asks a more sophisticated question, suggests possible bases for sample differences, and in the process hypothesizes reasons why the differences might be there.

2.9.7: Narrow Stimulus Sampling

Although the usual concern in generality of results has to do with sample characteristics and whether the findings extend across different subjects, equally relevant but less commonly discussed is the extent to which the results extend across the stimulus characteristics of the investigation (Wells & Penrod, 2011). The stimulus characteristics refer to features of the study with which the experimental manipulation or condition may be associated and include the experimenters, setting, interviewers, or other factors related to the experimental arrangement. Any of these features may restrict generality of the findings. That is, it is possible that the findings occurred under the very restricted or narrow stimulus conditions of the study.

There is a way in which you are already familiar with the problem. If I did a two-group study and used only one subject in each group, you would be likely to say, "Yes, but will the results generalize or apply to more than just that one subject who received the special condition?" Narrow stimulus sampling is very much the same question and goes like this, "yes, but will the results generalize to more than one stimulus condition provided in the study?"

The most common occurrence in psychological research pertains to restricted features of the experimenters or materials used in the study itself. For example, use of one experimenter, one therapist, one taped story, or vignette presented on a laptop or the Web may restrict generality of the results. Something about those stimulus conditions may be unique or special. Will the results show beyond this very restricted set of stimulus conditions?

Here is the case where it is difficult to have a plausible rival hypothesis as to why the stimulus conditions may be contributing to the results. Early in the development of a new intervention, often a single investigator may conduct a very well-controlled study. Let us say, patients are assigned to a new and improved treatment, traditional treatment, or wait (no treatment). The investigator

developed a new treatment and is the logical choice to serve as therapist. Let us say she administers all conditions. At the end of the study, patients in the investigator's special treatment did better than patients in the other groups. A potential problem as a threat to external validity is that perhaps the results would not generalize to other therapists. Something about the investigator administering her favorite treatment (more enthusiasm, persuasive appeal) may make the results specific to her. We could protect against that prospect and test this directly by including at least two therapists administering each of the conditions. Then at the end of the study, we could test whether the effects (conclusions) varied as a function of who administered the treatments. That would be an excellent addition to the study.

> The setting and restricted setting can serve as a basis of narrow stimulus sampling and raise questions about generality of the results.

Reviews of psychotherapy research consistently conclude that treatment is effective for a variety of clinical problems (e.g., Nathan & Gorman, 2015; Weisz & Kazdin, 2010). A perennial concern is whether the effectiveness of treatment in well-controlled studies is greater than the effects obtained in clinics outside of the laboratory.[5] Clearly, generality of results from controlled clinic/laboratory settings to "real-life" clinics raises a critical issue. Among the things we know is that once one leaves the controlled setting, the treatments tend not to be administered with the same care (e.g., diluted versions) and probably that is one reason why effects do not transfer (see Weisz, Ng, & Bearman, 2014). Also, many differences in the clients (who is seen in treatment, what types of problems they present), in who provides treatment, and in how carefully treatment integrity is monitored could explain setting differences. If these were controlled, setting may or may not make a difference in the effects of psychotherapy.

Concern with generality across settings extends to medical treatments too. Treatment and prevention of HIV/AIDS for example begin with controlled trials where the intervention (e.g., a "cocktail" combination of drugs–antiretroviral therapy) can be closely monitored and participants are recruited especially for the project. Once the effect is demonstrated, the treatment is extended widely to other circumstances where administration, delivery, and diversity of people are less well controlled. A recent demonstration showed that the results did indeed generalize from a controlled to community setting (Tanser, Bärnighausen, Grapsa, Zaidi, & Newell, 2013), but this is not always the case. As I mentioned, a highly controlled trial can show what can happen (in principle, when all is controlled) and then we extend this to natural settings where such controls and conditions are not easily implemented to if what does happen is similar to what was shown in research.

2.9.8: Additional Information on Narrow Stimulus Sampling

An elegant intervention study is to evaluate the impact of the treatment across two or more different types of settings in the same study. This is only occasionally done because of the logistics of implementing interventions in even one type of setting. Yet, as an example, in a large-scale treatment study, 24 public and private clinics (including >2,700 individuals with depression or anxiety) were assigned randomly to a stepped-care intervention (psychoeducation and then if needed interpersonal psychotherapy) administered by lay counselors or treatment as usual (regular clinical care) as administered by trained health care workers (Patel et al., 2010). Medication was available as was specialist attention (health professional) for suicidal patients. At 6 and 12 months after treatment, the intervention group had higher rates of recovery than a treatment-as-usual control group administered by a primary health care worker, lower severity symptom scores, lower disability, fewer planned or attempted suicides, and fewer days of lost work (Patel et al., 2010, 2011). Interestingly, the setting mattered. The results were quite strong in the public clinics but not evident in the private clinic. Among the possible reasons is that those seen in private clinics were receiving good quality care already in the treatment-as-usual control condition and therefore differences between the two treatment conditions may have been less easily detected.

Returning to laboratory experiments, usually it is not easy to pose a rival interpretation as to why one stimulus condition (experimenter) could restrict generality of the results. Yet, it is usually a better study if you can take that question of restricted sampling of the conditions presented to the subjects off the table. And stimulus sampling need not be large. Instead of one experimenter use two or more or instead of one vignette to present the manipulation or experimental condition, use two or more.

The goal: At the end of the study, we would like to say that the experimental manipulation did not depend on special features and ancillary features (something about the experimenter or vignette) of the study. Alternatively, if the results did depend on one set of characteristics rather than another, that would be very important to know.

As a summary statement, restricted stimulus conditions can be a threat to external validity. The narrow range of stimuli may contribute to the findings, and the results may not extend beyond these stimuli conditions. The implications of including a narrow range of stimuli in an investigation extend beyond a threat to external validity. We will address the issue again in the context of construct validity. At this point, it is important to note that the stimulus conditions of the experiment and settings in which research is conducted may very much relate to and hence limit generality of the results.

2.10: Additional Threats to External Validity

2.10 Classify each of the additional threats to external validity

Reactivity of experimental arrangements, reactivity of assessment, test sensitization, multiple-treatment interference, novelty effects, and generality across measures, setting, and time comprise the balance of major threats to external validity.

2.10.1: Reactivity of Experimental Arrangements

The results of an experiment may be influenced by the fact that subjects know they are being studied or that the purpose is to examine a particular relationship between variables.

> Reactivity refers to being aware of participating and responding differently as a result.

The external validity question is whether the results would be obtained if subjects were not aware that they were being studied.

In a vast majority of psychology studies, subjects (e.g., college students, recruited subjects from the Web) must be made aware that they are participating in an investigation and provide informed consent that they understand the purposes and procedures of the study. Investigators often try to obscure the focus or purpose of the study (e.g., "memory and emotion experiment" when the purpose is seeing if false memories can be induced). Yet, this does not alter reactivity per se. The issue is whether participants know they are participating in an experiment. There are exceptions where subjects do not know: those studies that examine records (e.g., medical or police records, school files) that are routinely available in a setting and where the individual subject cannot be identified. In these studies, informed consent may not be required and subjects then will not be told that they are in a study.

> If virtually all experiments are likely to be arranged so that subjects are aware that they participating in a study, why even discuss reactivity? The answer is that we want to reveal relations that are not necessarily restricted to the particular settings in which the relations are demonstrated.

For many types of experiments (e.g., nonhuman animal studies with rodents), we can be assured that reactivity is not likely to be a problem. We can assure this further by having no observers (humans around) and collect data automatically (machine) and observed by camera. But for human studies in laboratory paradigms, participants have their own motivation and interpretation of what is transpiring. If they are aware of participation, the effect of the experimental manipulation can be affected by these motives and interpretations. This is a threat to external validity because the generality question is, would these results be obtained if the subjects were not aware that they were participating in a study? Some situations might clearly raise the question. For example, in a study of altruism, sharing, kindness, prejudice, or table manners, perhaps people are a little more giving and "well behaved" than they would otherwise be if they did not know they were being watched in an experiment. In some cases, experimental situations can be set up where cameras record behavior while the individuals are left to believe that they are on their own. Or if they are informed that cameras are present, over time this may be less reactive than the situation in which a live experimenter was present.

Reactivity can be a matter of degree. Just because the subject is aware of participating in an experiment does not mean that the results will be altered. Subjects can vary in the extent to which their performance is altered and experimental arrangements can vary in the extent to which they are likely to foster special reactions. Even so, as one designs an experiment or a set of experiments, it is valuable to see whether key relations of interest are evident when the subjects are aware or not aware that they are participating.

Reactivity has an interesting twist in light of psychological research on unconscious process. Awareness of being involved in an experiment usually refers to the subject being able to state, realize, know, understand, and reply when asked about whether he or she is participating in an experiment. Yet, we often respond to cues that are out of our conscious awareness (Bargh & Morsella, 2008). Cues in the environment prime (or promote) thoughts and actions even though they are out of awareness. For example, some ancillary visual cue in the setting (e.g., a brief case) or some background smell (e.g., all-purpose cleaner) influences performance on experimental tasks (e.g., increases competitiveness and being neater, respectively) (see Thaler & Sunstein, 2008). When subjects are asked why they performed in this or that way, they cannot label the experimentally manipulated background cues that led to their behavior. Related, we have known for some time that the sight or presence of a gun can increase aggressive thoughts (e.g., Anderson, Benjamin, & Bartholow, 1998; Bartholow, Anderson, Carnagey, & Benjamin Jr, 2005). Again, one does not have to be aware of the cues consciously for them to exert influence. For present purposes, reactive arrangement can occur whether or not the subjects can explicitly state they are in an experiment.

The external validity question is whether there is something about the arrangement in addition to the experimental manipulation that may have contributed to the results. And, would the results be evident without that "something"?

2.10.2: Reactivity of Assessment

Reactivity of assessment is distinguishable and focuses on the measures and measurement procedures.

In most psychology experiments, subjects are fully aware that some facet of their functioning is assessed.

- *If subjects are aware that their performance is being assessed, the measures are said to be* ***obtrusive.*** Obtrusive measures are of concern in relation to external validity because awareness that performance is being assessed can alter performance from what it would otherwise be.
- *If awareness of assessment leads persons to respond differently from how they would usually respond, the measures are said to be* ***reactive****.*

In research, the fact that clients are aware of the assessment procedures raises an important question about the generality of the findings.

Will the results be evident (generalize) to other measures of which the subject was unaware?

What do you think?

The heavy reliance on self-report and paper-and-pencil inventories conveys the potential problem. These measures are among the most types amenable to subject distortion and bias. One can convey an impression or censor responses more readily on such measures. There are well-studied response styles or ways that individuals approach completing measures including the tendency to say yes and to present oneself in a socially desirable light. More will be said of these later. Obviously, these are more likely to influence the results when subjects are aware that their behavior or other responses are being assessed. The external validity question is whether the results would be obtained if subjects were unaware of the assessment procedures.

2.10.3: Main Strategy for Combatting Reactivity

The main strategy to combat reactivity and obtrusiveness of assessment is to include in a study a measure where the purposes are not so clear and where distortion is less likely.

For example, in a controlled study designed to prevent child abuse, mothers who used physical punishment and were high in anger toward their children participated in an intervention or no-intervention control group (Peterson, Tremblay, Ewigman, & Popkey, 2002). Parents kept daily diaries that just asked open questions about the children and how the parents responded. No questions were asked about physical punishment or the use of new practices (time out, ignoring misbehavior) that had been trained. The results supported the effectiveness of the intervention. The assessments may have been minimally reactive because they did not specifically ask about harsh practices that individuals may have been more reticent to acknowledge.

Another type of measure focuses on implicit attitudes or views and consists of individuals responding to a laboratory task where stimuli are presented perhaps along with positive and negative words. A measure derived from this is one's implicit attitude about the topic of phenomenon. For example, in one study newly married couples were assessed in relation to marital satisfaction at the beginning of the study (McNulty, Olson, Meltzer, & Shaffer, 2013). A self-report questionnaire was to measure satisfaction at this initial point in the study. In addition, an implicit measure was used where each person responded to a lab task of associating positive and negative words with a photo of their partners as well as with other individuals. From the measure, one could obtain an implicit (i.e., not explicit) view of one's partner and the positive or negative valence. Interestingly, the results indicated that self-report measure did not predict marital satisfaction over the course of the study. The implicit attitude measure did. Interestingly too is that the self-report and implicit measure did not correlate with each other. The point here is that there are measures that can identify out of consciousness views or reactions, and these are likely to be less vulnerable to reactivity.

There are many other measures (e.g., of biological processes and reactions, eye tracking, reaction time) where one knows that measurement is going on but the measures are less amenable to distortion on the basis of motivation or concerns of participants. In designing a study, it is usually advisable to use measures with different methodologies (e.g., not all self-report, not all direct observation). Among the reasons is to ensure that any finding is not restricted to a specific method of assessment. This is very much like the concern about narrow stimulus sampling. The reactivity issue adds another component. If it is possible to include measures that are less transparent or reactive, then usually that is an excellent addition to the study.

At this point, it is important to note that reactivity comes in two major forms that can influence generality of the results:

1. The *experimental arrangement*, discussed previously, and it refers to whether subjects believe they are participating in an experiment or might pick up cues that guide their behavior whether or not they are aware.
2. The *assessment procedures* and whether the subjects are aware of being assessed and presumably can alter their performance as a result.

In any given study, reactivity may or may not be of interest. More broadly in psychological science, we

certainly want to establish findings that are not restricted to when subjects are aware they are in an experiment and being assessed.

Critical Thinking Question

What is the main strategy that you can use to combat reactivity and obtrusiveness of assessment? Recall the example of the controlled study designed to prevent child abuse. Why would a study such as this work well? Come up with your own idea for a measure where the purposes are not so clear and where distortion is less likely.

2.10.4: Test Sensitization

In many investigations, pretests are administered routinely. These refer to assessment at the beginning of a study in advance of whatever the experimental manipulation or intervention will be. The purpose is to measure the subject standing on a particular variable (e.g., reading skills, anxiety). There are many methodological and statistical benefits in using a pretest.

At the same time, *administration of the pretest may in some way sensitize the subjects so that they are affected differently by the intervention, a phenomenon referred to as pretest sensitization*. Individuals who are pretested might be more or less amenable or responsive to an intervention (e.g., treatment, persuasive message) than individuals who are not exposed to a pretest merely because of the initial assessment.

> Essentially, a pretest may alert someone to attend to some manipulation or event in a different way from what they would have without a pretest.

This does not have to be a conscious process where the individual makes the connection between the assessment and the experimental manipulation or says "aha" and has some special insight.

As an example, consider an investigator who wishes to examine people's views toward violence. The hypothesis may be that viewing violent movies leads to an increase in aggressive thoughts and a greater likelihood of engaging in aggressive acts. The investigator may wish to evaluate views of people after they see a violent gangster film. The film may be shown at a movie theater, a laboratory setting where volunteers view the movie on a laptop, or video segments played online for subjects recruited through the Internet. Let us take the movie theater as the example. As patrons enter the lobby, they complete the measure right before viewing the film, and then on their way out of the theater, they complete the measure again. They are promised a coupon for a small bag of popcorn (probably the equivalent of U.S. $20) for their next visit once they complete both measures.

Here is how pretest sensitization can fit in.

It is possible that administration of a test before the film is shown (i.e., the pretest) makes people view and react to the film somewhat differently from their usual reaction. Perhaps the questions heighten sensitivity to certain types of issues or to the entire topic of violence, which may not have otherwise been raised. At posttest performance, how subjects respond is not merely a function of seeing the movie but also may be due in part to the initial pretest sensitization. Hence, a possible threat to external validity is that the results may not generalize to subjects who have not received a pretest. Administering a pretest does not necessarily restrict generality of the results. It does, however, raise the question of whether non-pretested individuals, usually the population of interest, would respond to the intervention in the same way as the pretested population. As with other threats to external validity, there is no challenge here about the finding itself, but there is a question of whether the relationship between viewing a violent movie and views of violence would be evident without some sensitization experience that immediately precedes the movie.

A potential confusion is useful to address here. Earlier I discussed testing as a threat to internal validity. That refers to improved (or could be worse) scores merely as a function of completing the measure more than once. Now we have discussed pretest sensitization. This is the possibility that the pretest plus the intervention lead to a special outcome at posttest. The intervention might not show that same effect if it were not for some pretest that helped increase the impact of the intervention. This is not the effect of repeated testing but the effect of special impact of the pretest.

I mentioned an example previously designed to prevent sexual assault among military personnel. Individuals received a special training program to increase their knowledge, empathy, understanding of military rules and policies, and women's perspectives related to sexual assault or received no intervention (Rau et al., 2011).[6] An interesting feature of the study was an effort to evaluate testing and pretest sensitization. So some subjects received only the posttest and others both pretest and posttest. The results: there was a testing effect so that even without a special intervention military personnel improved (e.g., became more sensitive, empathic, etc.) on the measures. But there was no pretest sensitization effect. This means that individuals who received the intervention improved equally well (and better than no treatment) whether or not they had the pretest.[7]

2.10.5: Multiple-Treatment Interference

In some experimental designs, subjects are exposed to more than one experimental condition. This could be an experiment manipulation, sequence of tasks, or two or

more interventions. Perhaps subjects are asked to evaluate different stimulus materials (e.g., faces, stories, videos) that vary in some way to test a critical hypothesis. When more than one task is presented, it is possible that performance on the second or third task is influenced by the preceding history of a prior task.

> Multiple-treatment interference refers to drawing conclusions about a given manipulation or intervention when it is evaluated in the context of other manipulations. (Although tradition uses the term "treatment," the threat refers to instances in which subjects receive more than one condition in an experiment; that does not have to be "treatment.")

This is an external validity issue because the conclusion drawn about one intervention may be restricted (and not generalize) to those circumstances in which prior manipulation or intervention was not provided. Stated another way, the effects obtained in the experiment may be due in part to the context or series of conditions in which it was presented.

As an illustration, assume an intervention is used for individuals who were in a depression program where they received cognitive behavior therapy. Let us say further that many individuals who did not respond now were subjected to another form of therapy (e.g., interpersonal psychotherapy). Let us say further that most of the individuals did respond (recover from depression) after this second treatment. It is possible that interpersonal psychotherapy worked in the context of having a prior treatment first. Perhaps interpersonal psychotherapy would have not been effective or as effective without the sequence of cognitive behavior therapy followed by interpersonal psychotherapy. Just because the first treatment did not seem to be effective does not mean that it had no effect. It may have sensitized or somehow increased responsiveness to the second intervention. That would be an example of multiple-treatment interference, namely, the effects of the second treatment may not generalize to situations in which that treatment is presented on its own.

In treatment research (e.g., for psychotherapy or medication) occasionally cross-over designs are used in which some individuals receive intervention A (e.g., for a week) followed by intervention B. Other individuals receive the same interventions but B first and then A. After each treatment, measures are administered to evaluate outcome. Whatever the outcome, the effects of B might be due to B all by itself or be due to B only when preceded by A. We cannot tell. Adding another group that receives the same treatments in reverse order can tell. Is B equally effective when it is first or second in the sequence? In such designs, one can evaluate whether a given treatment (A or B) has different effects with and without potential interference of a prior treatment.

2.10.6: Novelty Effects

As a threat to external validity, novelty effects refer to the possibility that the effects of an experimental manipulation may in part depend upon their innovativeness or novelty in the situation.

> The effects of the intervention may depend upon and be limited to the context in which it is administered.

The context may make the experimental manipulation salient or otherwise novel in some way. Consider an example of novelty effects well outside of psychology. In the United States, thousands of motor vehicle accidents (e.g., ~30,000) occur each year between fire trucks and other vehicles and are the second highest cause of death (after overexertion/stress as the number one cause) among firefighters (Donoughe, Whitestone, & Gabler, 2012). Some evidence has suggested that yellow (or lime/yellow) fire trucks, compared to more traditional red fire trucks, have significantly fewer accidents with cars, and when they have accidents they are less serious (http://www.apa.org/research/action/lime.aspx). The usual interpretation of this finding is that the human eye has greater difficulty in perceiving red relative to many other colors, including yellow. Because drivers more readily see yellow trucks, they can avoid the trucks more easily and hence fewer accidents are likely.

> Consider the role novelty might play as a threat to external validity. Quite possibly, the reduced accident rates associated with yellow fire trucks are due in part to the fact that such trucks are quite novel. That is, they depart from the vast majority of red trucks that still dominate in the United States. The reduced accident effect could be restricted to the *novelty* of introducing yellow trucks and not the color itself.

Perhaps if most fire trucks were yellow, the benefits would be lost. Indeed, against a sea of yellow trucks, red trucks might be associated with reduced accidents because of *their* novelty. This is not a trivial point because loss of life is involved and we want to know precisely what to do. If the color of fire trucks were the critical issue, then changing to yellow trucks would be the obvious strategy. If novelty accounts for the effect, this leads to a different strategy. One would encourage changes in fire trucks. Whenever trucks are replaced or repainted, perhaps their color ought to change and the color (from red to yellow and perhaps back to red or some other color) might not be too important. Or if it is novelty, patterns on the trucks or light displays that could be varied to make them novel might be called for. The research may sound silly but without understanding why reduced accidents occur, it is easy to adopt a strategy that will have little impact. For example, a change to all yellow or mostly yellow fire trucks may have no enduring impact on accidents if novelty is responsible for

the effect. (My master's thesis on stealth and camouflage fire trucks with sirens pitched so only dogs could hear them shed interesting light on this matter but that story is for another time.) The external validity issue is this. The intervention (yellow trucks) may be restricted to situations in which that intervention is novel.

In passing, it is worth noting that most fire trucks in the United States remain red and many fire departments that tried yellow trucks are returning to red (Thompson, 2010). Among the reasons is that the public has strong recognition of fire trucks as being red. Also, technology has improved so that fire trucks can have more reflective materials and make themselves conspicuous in other ways (sounds and sights) than changing their color. All that said, we do not seem to have a verdict with solid data one whether any particular color now decreases accidents and fatalities and if it does whether it is novelty (changing colors) or the color.

The presence of novelty effects is difficult to evaluate. The situation in which it is likely to emerge is when some intervention or manipulation is compared to no treatment or no manipulation. The study seems controlled, but the intervention or manipulation may have worked because of its novelty that in fact was not controlled. A new intervention (e.g., educational program, diet, treatment) when first introduced may seem to be effective because of its special procedure effects or because of its novelty (or some combination). Effects due to novelty would wear off over time as the intervention is extended and becomes the norm rather than novelty. It may be that we must couch new interventions as "new and improved," very much like the latest in smartphones, smartwatches, ear buds, headsets, soaps, cereals, automobiles, and shampoos we purchase, and they will be more effective in part because of their novelty. When a new intervention is introduced, we may not know whether it is the "new" (novelty effect) or the "improved" (putatively better procedures or techniques) that accounts for the change. In such cases, novelty becomes a threat to external validity; that is, the effect may not generalize to situations in which it is not a novelty.

2.10.7: Generality across Measures, Setting, and Time

This is a catch-all category of threats to external validity. The potential threat to consider is whether there is any facet of the measure, setting in which the study was done, or time that might restrict generalization of the results.

- As for measures, were the results obtained on measures unique to those particular measures rather than to carry over to other indices of the construct? For example, in treatment studies, individuals may decrease in depression on standardized psychological measures (paper and pencil, interviews).
- Do the intervention effects carry over to other indices of how individuals are doing in everyday life?
- Or when voters are surveyed (self-report) about for whom they will vote, to what extent does that generalize to other measures (their actual voting behavior)?

For setting, the study was conducted under special circumstances (e.g., some lab arrangement).

- *To what extent are the findings likely to be restricted to that setting?* For example, many prevention programs are first evaluated in the schools. Yet, there may be many features of the schools (which ones were selected, in what neighborhoods, under what circumstances) that optimize the likelihood of obtaining an effect of the program.
- *Will the results generalize to other settings that perhaps have more impediments (e.g., financial, administrative) that could limit the effects of the program?*
- *Finally, are the findings restricted to a particular point in time?* There are different variations. The first pertains to timing of the assessment. At the end of the study, assessments are administered and groups may be different.
- *Would these effects be evident one month or one year later?* In most experiments (e.g., in labs with college students), this is not necessarily of interest. The study is done, data are collected, and all the conclusions one wants to reach are for that moment in time. In other cases where there are interventions (education, treatment, prevention, counseling), we may care about immediate change but also more enduring effects.

  An external validity question that can be raised is whether the same results would have been obtained had measurements been taken at another time, say, several months later.

For example, in the context of psychotherapy research, the effectiveness of treatment usually is evaluated immediately after the last therapy session (posttreatment assessment). Yet, treatment studies occasionally show that the conclusions reached about a particular treatment or the relative effectiveness of different treatments can vary from posttreatment to follow-up assessment (see Kazdin, 2000). In some cases, treatments are no different at posttreatment but are different at follow-up; in other cases, treatments are different at posttreatment but no different at follow-up. Most psychotherapy studies do not include follow-up assessment, so we do not know how pervasive these variations are. But the general point for external validity is not about psychotherapy studies. In short, conclusions about the effectiveness of an intervention may depend on when

the assessments are completed. Stated as a threat to validity, one might say, yes this intervention is more (or less) effective. However, would this conclusion apply to another time period (e.g., in a year from now)?

A second way in time can relate to external validity pertains to cohorts.

2.10.8: Cohorts

> A cohort refers to a particular group of people who have shared something over a particular time period.

This usually refers to a group born in a particular period or a group that is studied but gathered (assessed) at a particular point in time and followed. For example, a type of research we will discuss later is referred to as a birth-cohort study. This is a study in which all or most individuals born at a particular time are identified and participate in a study over several years, often decades. That is, the "cohort" or group initially identified at a particular time (e.g., all individuals born within a given year in a given city) is assessed at multiple points over time. It is possible that some relationship between variables would differ for different cohorts. That is, a finding obtained with one cohort might not generalize to another time period. This is easily envisioned when speaking about a given generation and their parents.

Society can make dramatic changes over time and the meaning, importance, and prevalence of many things change. For example, tattoos, use of marijuana, social contacts (networking), and age of first sexual activity have changed in meaning or use for individuals now in their 20s and 30s from what that was for individuals now in their 60s. That is, the cohorts have different experiences. In relation to research, timing of a finding might be an external validity concern if there was a reason to suspect that the finding depends on unique features of a cohort (group, generation) that do not apply to other cohorts. Psychology rarely studies cohort effects quite this way, but it is important to be aware of the possibility that a particular finding may not apply to people at different time periods. Perhaps some cohorts are more or less sensitive to a particular influence because of some other facet of the culture associated with their unique period of development.

Within a given cohort, a finding also might be restricted in time. For example, a given relation between two (or more) variables (sex and problem-solving skills; responsiveness to persuasive appeals and peer pressure, factors that influence attractiveness toward a potential partner) may be studied in a sample of college students in their early 20s.

Does the finding generalize to other periods in which the relation would be assessed for these same subjects? For example, if this sample or indeed another sample were assessed in their 40s, would the relationship of some variable and attractiveness be the same?

What do you think?

Common sense and experience convey that our beliefs, values, and views change over time due to historical events (e.g., getting married, having children) and maturation (e.g., gaining experience, learning, biological changes in hormones). Presumably, these would be reflected in the relations among variables assessed in a psychological experiment. The point is noting that time of measurement can affect conclusions quite broadly. Of course, describing and understanding changes in relations among variables over time is part of developmental psychology and life-span research. This is a case where the concern of external validity has important substantive (developmental) implications.

2.11: When We Do and Do Not Care about External Validity

2.11 Evaluate the idea of proof of concept

One might think that we always want to know the extent to which our results generalize beyond the specific experiment we have conducted. Indeed, what good would a finding be if it were restricted to just the situation in which we provided a test? Well it might be absolutely great! This brief discussion is important in relation to designing your own research as well as evaluating the research of others.

2.11.1: Proof of Concept (or Test of Principle)

A critically important type of work is designed to show if something can occur.

> Proof of concept is a test to see whether something could occur that is important in principle or in theory.

Usually a special situation is arranged, and that is likely to be artificial, contrived, and not very much like everyday experience at all. The purpose is to see if something can happen even if this is not related to how something does happen in the world. You know about this already. For example, particle physics seeks to understand the basis of matter, i.e., the underpinning of all things. And in the process, physicists look for theoretically predicted particles (e.g., Higgs-Boson), which are only evident in special circumstances (bizarre collisions of matter in special colliders of matter). At the end of a demonstration, we do not ask, "Yes, but can these particles be shown to appear in my home or how about California where everything seems to be different?" Nor do we say—this was only one collider (narrow stimulus sample) and who knows if the results would generalize to other colliders. There are scores of questions in natural, biological, and social sciences that are about proof of concept.

Within psychology, for example, we would like to know if experiences and memories of traumatic events can be eliminated or replaced. Traumatic memories can be lifelong and debilitating (e.g., posttraumatic stress disorder from experiences in war, rape, child abuse, exposure to violence). Could we show under any circumstance that a traumatic memory can be erased? Consider this demonstration. Precisely how long-term memories are coded and maintained rely on a particular protein and once that is "turned on" the memory remains. Researchers have exposed sea snails (*Aplysia*) to shock (to induce trauma), then showed that aversive reaction was maintained (remembered, as reflected by recoiling when touched after the experience of being shocked), and finally showed that the memory could be undone (erased, eliminated) by manipulating the protein that maintains the memory (Cai, Pearce, Chen, & Glanzman, 2011). The results? This chemical undoing returned the snails to responses as if they had not had the traumatic experience. No doubt once this finding for sea snails circulated on Facebook, the news went viral. What about us—humans? Of course, we want to know whether this finding could generalize to human traumatic memory, but it is not a criticism in basic research to attack that as a weakness of the study. Proofs of concept are just that—can something be identified or shown to occur? We want to know the basic question: can traumatic memory ever be erased in any instance? We first need demonstrations of the principle. Any generality issues are not part of the project at that early point. Also, this basic finding might suggest the mechanisms of action, that is, how traumatic memories are coded and erased and that is likely to have broad generality (e.g., across species).

More broadly, there are many questions that can guide research. It is useful to identify the broad purpose of a study you are about to do. Do I want to show some principle or concept whether or not that reflects how things really are? Another way of saying this, do I want to take some phenomenon of the "real world" (romantic love, trust, contagion, emotional regulation, love of methodology texts) and bring it into the laboratory to study some fine point that observation in the world does not so readily allow? It is not so easy to state that external validity or generalizability of a result is or is not important in an all-or-none fashion. It is important to note here that external validity is not always a concern in an experiment and proof of concept studies is the prime example. Proof of concept studies can advance our understanding enormously.

2.11.2: Additional Information on Proof of Concept

There is a sidelight to proof of concept research that is worth mentioning. Science increasingly is under public scrutiny as it should be given that the goals are the betterment of public life and the means to achieve these rely heavily on tax (federal, state) and private (foundations, donors) funding—all "public" funding in some way. I mention this because proof of concept research is an easy target (e.g., for U.S. Congress, public figures, news media, and the public at large) because it is not well understood.

As an example, a recent study showed that familiarity is related to sexual attraction and mate selection, what hormone and site in the brain appears to be involved, and how familiarity and lack of familiarity with various possible partners get translated into activation or inhibition of that hormone (Okuyama et al., 2014). Sounds good so far—but this was done in fish. This was a meticulous proof of concept study to isolate a process. It would be easy to ridicule the research by superficial analysis. Someone could say, "Why do we care about attraction in fish, even a fish (medaka) most people never heard of? What a waste of taxpayer dollars." There is a long history of precisely such criticisms and it continues. The fish example, from a scientific standpoint, is huge. If we can better understand attraction and underpinnings of social relations, the potential impact of this study and the line of work it spawns could be great. In many psychological dysfunctions, social relations are disrupted or dysfunctional (e.g., ASD, schizophrenia); for even a larger number social relations and social attraction also can interfere with adaptive functioning (e.g., individuals who are lonely or isolated from others).

Social relations relate to physical health too and understanding all facets of how they emerge could be important.

All my comments are speculative and also shortsighted. We cannot usually predict the long-term impact of proof of concept studies, but they account for our smartphones and computers, our prosthetic devices (e.g., artificial limbs) and control over these limbs, and so on. Proof of concept research actually is the core of science, and that includes psychological science (e.g., on such topics as prejudice, decision making, addiction). I am not saying all proof of concept is important and some research that seems silly might well be—I am not the arbiter of that. I merely mention this because proof of concept research with its extreme care and control of internal validity to show what might be possible is the greatest target for misunderstanding and fairly easy to mock without the understanding.

2.12: Managing Threats to External Validity

2.12 Examine the importance of determining the relevance of a threat to external validity before it is managed

As usual, the first step in managing threats is to consider each one individually and determine if it is relevant to the goals of the study. That is, precisely what are the goals of

the study in relation to generality? As already mentioned, for studies that are proofs of a concept, generality is not necessarily this initial concern. In other studies (e.g., lab studies with college students), generality may not be of a concern either. It is important to be clear about the purpose and select subjects and conditions consistent with that. It may well be that subjects are selected with a narrow set of characteristics (e.g., same sex, close in age, of one ethnicity) because that follows from the goal of the study. A key issue about sampling and external validity is how the investigator eventually discusses the findings. It is important to be careful in discussing the findings not to go beyond what can be said in relation to the conditions and subjects to whom the results might apply.

One cannot anticipate all of the external validity threats to be managed in a given study. Yet one can be alert to several issues and a few practices that address key threats we have mentioned. We have discussed several issues about the sample. Managing this concern has to do with conveying why the sample one is studying is a good one to use for the study. Two weak reasons are that the sample was easy to obtain and that other people have used the sample. A stronger reason might be that one wants to identify a particular relation and the study is to evaluate a proof of concept, i.e., can something be shown to occur.

We also discussed narrow stimulus sampling. In any instance in which stimuli are presented to the subjects or research assistants are interacting with subjects and could exert impact, include two (or more) examples. If one is presenting a vignette or case for subjects to react to, present two vignettes or cases that vary slightly characteristics. This will permit one to test and evaluate whether the findings were associated with one experimenter. This is an easy item to anticipate and to integrate in a study. In most cases, the vignette, case, experimenter, or other stimuli that have been sampled will not make a difference when tested statistically. Yet, it is reassuring to be able to say that and to take this threat off the table.

Reactivity of arrangements and reactivity of assessments have their limitations in what one can control. Informed consent usually requires that subjects are aware of all procedures and measures. Even so, one can make less salient the specific purposes of the study—the general focus can be described but nothing that would convey specific hypotheses. Thus, subjects may behave in a special way because they are aware but that awareness could not bias the direction of the hypotheses. Reactivity of assessments offers several better options. Measures are likely to vary in the extent to which awareness of measurement can influence the results or to do so in a biased fashion. For example, one might want to study the extent to which individuals smoke cigarettes, use marijuana, or engage in alcohol consumptions. In such studies, daily or weekly self-report measures may be used. Self-report is fine but as a method is quite amenable to distortion. Additional biological measures to measure blood levels, drug or alcohol use, for example provide indices less amenable to reactive influences. Also, test sensitization was noted as a threat to external validity. Here use of measures that are less likely to sensitize individuals to a later manipulation is useful. No measure is perfect, and using multiple measures with different strengths and liabilities is a wise strategy for a variety of reasons.

Some of the other threats such as multiple-treatment interference will be mentioned later. Most studies do not include multiple conditions for the same subject. That is, each subject is exposed to all of the conditions and assessment is made at multiple points to see if performance differs. In such circumstances, it is important to balance the order of conditions so that subjects do not receive just one order of the conditions.

2.12.1: General Comments

The threats to external validity only begin to enumerate those conditions that might restrict the generality of a finding. All of the conditions that are relevant to the generality of a finding cannot be specified in advance and that is why it is difficult to delineate general strategies to manage them. In principle, any characteristic of the experimenters, subjects, or accouterments of the investigation might later prove to be related to the results. If one of the threats applies, this means that some caution should be exercised in extending the results.

One way of conceiving many of the threats that were mentioned and others that might be raised is the notion of context. It might be that the experimental manipulation or intervention achieved its effects because of something about the context in which it was studied or demonstrated. If the intervention occurred after some other intervention (multiple-treatment interference), under arrangements in which subjects knew they were being studied (reactive arrangements), under conditions in which the experimental intervention might seem quite novel in light of one's usual experience (novelty), or with assessments that may be especially prone to show effects (reactive assessments, test sensitization, right after the manipulation but not months or years later), one could raise questions of external validity. The effects may be carefully demonstrated (internal validity was well handled), but perhaps the findings would not be obtained or obtained to the same extent without one of these contextual influences. The degree of caution in generalizing the results is a function of the extent to which special conditions of the study depart from those to which one would like to generalize *and* the plausibility that the specific condition of the experiment might influence generality.

One cannot simply discount the findings of a study as a very special case by merely noting that participants were pretested, that they were aware that they were participating in an experiment, or by identifying another characteristic of the experiment.

> Enumerating possible threats to external validity in principle is insufficient to challenge the findings.

The onus is on the investigator who conducts the study to clarify the conditions to which he or she wishes to generalize and to convey how the conditions of the experiment represent these conditions. The onus on those skeptical of how well this has been achieved is to describe explicitly how a particular threat to external validity would operate and would be plausible as a restriction of the findings.

An example where I believe a plausible case can be made for a threat to external validity comes from a remarkable longitudinal study designed to test if a daily multiple vitamin reduced the rates of heart disease and stroke in years later (Sesso et al., 2012). More than 14,000 physicians who were at least 50 years old were assigned randomly to a daily multivitamin or placebo control condition and followed for an average of approximately 11–13 years. Main results—there was no difference between vitamin and placebo groups in the rate of heart disease, stroke, or death during the study period. Internal validity and many other features of the study are truly exemplary. Yet, for me, there is a lingering threat to external validity because of the sample. To me it is plausible to say that the results might not generalize too many other populations. The reason is that in the world doctors as a group compared to everyone else probably have better diets and habits (e.g., less junk food, less cigarette smoking that can deplete vitamins) and have generally high socioeconomic and occupational standing (e.g., better care for their daily health). Vitamins might be expected to make little difference in a relatively healthy group.

2.12.2: More General Comments on Managing Threats

It seems plausible (to me) that these findings might not hold true for people in developing and developed countries whose diets are not as healthful, who do not have access to suitable health care, or indeed who are below the poverty line and do not eat very much. In this example, one cannot say merely that these were doctors and the results may not generalize. One needs to explain why and the "why" needs to be plausible. You can judge whether my view is plausible.

Many conditions might be ruled out as threats to external validity on seemingly commonsense grounds (e.g., hair or eye color of the experimenter, season of the year, birth weight of participants, time of day the study was conducted). The reason is that it is difficult in most instances to come up with a theory that would convey how hair or eye color would influence or interact with the experimental manipulation to produce a special result that would not generalize to other conditions. Yet, in any given area unpredictable influences may be important. For example, I mentioned earlier how the likelihood of being granted parole is related to whether the parole board is hungry or had a recent break (Danziger et al., 2011). The overall point, seeming trivial details might well influence generality of a finding.

The task of us as consumer of research (e.g., students and other professionals, lay persons) is to provide a plausible account of why the generality of the findings may be limited. Only further investigation can attest to whether the potential threats to external validity actually limit generality and truly make a theoretical or practical difference. Of course, there is no more persuasive demonstration than several studies conducted together in which similar findings are obtained with some consistency across various types of subjects (e.g., patients, college students), settings (e.g., university laboratory, clinic, community), and other domains (e.g., different researchers, countries). Replication of research findings is so important to ensure that findings from an initial study are not likely to be due to various threats to internal validity or to chance. Replication is also important for external validity because further studies after the original one are likely to vary some conditions (e.g., geographical local, investigator, and type of subject) that extend the generality of the findings.

2.13: Perspectives on Internal and External Validity

2.13 Analyze the similarities and differences between internal validity and external validity

Internal and external validity convey critical features of the logic of scientific research. Internal validity is addressed by experimental arrangements that help rule out or make implausible factors that could explain the effects we wish to attribute to the intervention. Everyday life is replete with "demonstrations" that do not control basic threats to internal validity. For example, almost any intervention that one applies to oneself or a group can appear to "cure" the common cold. Consuming virtually any vitamin or potion from assorted animal parts, reading highly arousing material (e.g., on methodology and research design of course), texting a friend for 5 minutes, or playing with Facebook each day in a few days will be followed by a remission of cold symptoms. Pre and post assessments with one of the above interventions would no doubt show a reduction in cold symptoms (e.g., on a checklist of cold symptoms), improved

feelings of well-being, and changes in various biological indices (e.g., body temperature, swelling of the sinuses).

Did our intervention lead to improvement? Probably not. We can muse at the example because we know that colds usually remit without the above interventions. The example is relevant because maturation (immunological and recuperative processes within the individual) is a threat to internal validity and can readily account for changes. For areas we do not understand as well and where the determinants and course are less clear, a host of threats can compete with the variable of interest in accounting for change. Control of threats to internal validity becomes essential.

2.13.1: Parsimony and Plausibility

Parsimony and plausibility are quite pertinent to the threats to validity. Key threats to internal validity (history, maturation, repeated testing) are threats in part because they often are parsimonious interpretations of the data and often as or more plausible than the interpretation proposed by an investigator. For example, consider a study in which all subjects with a particular problem receive psychotherapy and improve significantly from pre- to post-treatment assessment. The investigator claims that therapy was effective and then foists upon on all sorts of explanations that mention the latest in cognitions, family processes, and brain functions to explain treatment (done rather well, I might add, in the second Discussion section of my dissertation). However, basic threats to internal validity are quite plausible as the basis for the change and are as or more parsimonious than an investigator's interpretation.

> History, maturation, and the other internal validity threats show broad generality across many areas of research and hence can account for many findings beyond those obtained in this particular study.

Consequently, as a matter of principle, the scientific community adopts the threats to internal validity as the more likely basis for explaining the findings if these threats are not addressed specifically by the design. This does not mean history, maturation, and so on *did* cause the changes, but it does mean there is no reason to resort to some specific explanation of the changes, when we have as plausible alternatives changes that can be pervasive across many areas of study.

If the investigator has in the study a control group and subjects were assigned randomly to some intervention or control conditions, then history, maturation, and the other threats are no longer very plausible or parsimonious. The skeptic must pose how history or maturation applied to one group rather than another (selection $\times$ history, selection $\times$ maturation) to explain the differential changes. This is often possible, but rarely parsimonious or plausible. The simpler explanation of the finding is that the experimental manipulation was responsible for the differences between groups.

Parsimony applies equally to threats to external validity. It is not reasonable to look at a finding and say in a knee jerk way, "Yes, but does the finding apply to this or that ethnic group, older or younger people, people of a different gender, or subjects without a pretest?" Parsimony begins with the assumption that most humans are alike on a particular process. This does not mean that most humans *are* alike on any particular process or characteristic. Parsimony is a point of departure for developing interpretations and not an account of the world. Absent evidence or theory, one does not merely propose differences. One needs a reason to suggest that generality is restricted beyond merely noting that the study did not sample all conceivable populations or circumstances in which this independent variable could be tested. There *are* restrictions in findings, and a given finding does not invariably generalize. That said, this does not mean all findings are suspect or limited unless broad generality is shown. Parsimony moves us in a more sympathetic direction, namely, the finding is the best statement of the relationship unless there are clear reasons to suspect otherwise.

2.13.2: Priority of Internal Validity

As a priority, the internal validity of an experiment usually is more important, or at least logically prior in importance, than external validity. One must first have an unambiguous finding before one can ask about its generality. Given the priority of internal validity, initial considerations of an investigation pertain to devising those conditions that will facilitate demonstrating the relation between the independent and dependent variables.

By emphasizing internal validity, there is no intention to slight external validity. For research with applied implications, as is often the case in clinical psychology, counseling, education, and medicine, external validity is particularly important. A well-conducted study with a high degree of internal validity may show what *can* happen when the experiment is arranged in a particular way. Yet, it is quite a different matter to show that the intervention would have this effect or in fact does operate this way outside of the experimental situation. For example, medications as well as vaginal gels have been used to treat and prevent HIV/AIDS. In studies, the goal is to evaluate use of the medication in controlled trials to evaluate the impact. Treatment and prevention are related here because having individuals with HIV/AIDS use effective practices not only maintains their own health but also decreases the likelihood that their partners will contract HIV. In controlled trials, treatment administration is carefully monitored and overseen to see if under the best conditions the interventions are effective. Once such trials are completed, interventions are often extended to the "real world," to see if the effects generalize to circumstances where the

intervention may not be so easily monitored and controlled. In some cases, treatment shown to be effective can be effectively extended to larger scale applications (e.g., Tanser et al., 2013), but in other cases the treatment does not work (Cohen, 2013). Real-world extension can raise problems such as drawing on a more diverse set of patients than those included in the original controlled trial and getting patients to follow the prescribed treatment (e.g., take pill, use vaginal gels).

When a finding does not generalize, the novice skeptic says, "What good is it to find something that works in a controlled setting if it does not extend to the 'real world?'" As mentioned in the discussion of proof of concept, initial studies are designed to see what can happen.

> In relation to the HIV/AIDS example, can we effectively treat or prevent the disease? For this the priority is internal validity to see if we can in principle achieve the change. This research might be conducted in very highly controlled studies and studies with human and nonhuman animals. At this point, internal validity rules.

Once we know we can, we become more interested in external validity. For example, we understand what happens with the treatment (e.g., mice or monkeys, or humans who can be very closely monitored and evaluated to be sure that treatment is being carried out). Now external validity becomes the priority. Can the intervention be extended to humans, to rural areas with few health care workers, to people who have multiple problems beyond just the condition we are trying to treat or prevent, and so on? Controlled studies are needed here too.

2.13.3: Further Considerations Regarding Priority of Internal Validity

The priorities of internal and external validity are a source of frustration when the research somehow relates to a significant psychological, medical, or environmental problem. It is easy to be frustrated with basic research that focuses on theory, animal models, and laboratory conditions. Obviously we want help for a particular problem now (e.g., schizophrenia in adolescents, Parkinson's or Alzheimer's disease in the elderly, cancers at all ages). Then we learn of research that "shows promise" and new insights in some laboratory under some esoteric condition in animals we are now quite sure we can even picture (sea snail, voles, zebra fish—sound like names for rock bands). The investigator is interviewed and says, of course this demonstration is promising but we are years away from application. Stated in terms of this chapter—the scientist is saying, "we have demonstrated the phenomenon and partially isolated the influences" (this is internal validity) but we are years away from knowing whether it will carry over to the circumstances we all care about (this is external validity). Yet, this is the order of research one often has to follow. Testing things directly in the world as a beginning point can slow the process because there are so many influences that can undermine showing an effect. We might discard effective interventions that could have been understood in controlled settings (with emphasis on internal validity), developed further, and then extended to the world (external validity). Also, highly controlled and indeed "artificial" circumstances need to be created to observe something that cannot be examined in nature. Subtle parenting practices, sources of stress, interpersonal interaction, opportunities for prejudice—all psychological processes that should be understood and then where implications exist extended to practice.

As one reads findings of research or plans a career in research, it is important to keep in mind that some of the best, most helpful, and practical research that has real-world impact begins with understanding the phenomenon of interest and studies that may have very little external validity. This is true in psychology but of course other sciences more generally. In clinical psychology, effective treatments have come from "curing" anxiety artificially induced in dogs and cats; in physics, understanding properties of laser light (electromagnetic radiation) has completely altered surgery where lasers are used instead of scalpels to cut. Basic research with careful attention to internal validity and with efforts to isolate and understand processes is critical.

> The goal of research is to understand phenomena of interest. Internal validity obviously is relevant because it pertains to many potential sources of influence, bias, artifact that can interfere directly with the inferences drawn as to why a finding was obtained. In the context of understanding phenomena, external validity has a very important role that goes beyond merely asking, "Yes, but do the findings generalize to other . . . (people, places, settings, and so on)?" When findings do not generalize, there is a very special opportunity to understand the phenomenon of interest.

Failure to generalize raises questions of "why." In the process, a deeper level of understanding of the phenomenon of interest is possible. It may be that the relation depends on the presence of some third variable (e.g., personality, sex, education of the subjects) or some artifact in the experiment (e.g., a threat to internal validity). Either way, establishing when the finding does and does not hold can be a conceptual advance.

Issues of external validity sometimes emerge, or at least ought to, when there is a failure to replicate a finding. Failures to replicate a finding sometimes are viewed as reasons to be suspicious about the original finding. Either the original finding was an artifact (due to threats to internal validity or other biases we will consider later) or

the finding was veridical but restricted to very narrow conditions of the original investigation (limited external validity). It is possible that some other variable provides the boundary conditions under which a finding can be obtained. Theory and research about that variable (or variables) can promote highly valuable and sophisticated research. As researchers, we often search for or believe we are searching for general principles that have widespread, if not universal, and intergalactic, generality. Yet, the value of a finding does not necessarily derive from its generality. Knowledge of a phenomenon entails identifying the conditions under which the findings may not apply and the reasons for seeming exceptions to what we thought to be a general rule. External validity issues are not mere afterthoughts about whether the findings "generalize," but get at the core of why we do research at all.

Summary and Conclusions: Internal and External Validity

The purpose of research is to understand phenomena peculiar to a discipline or specific area of study. This translates concretely to the investigation of relations between independent and dependent variables. The value of research derives from its capacity to simplify the situation in which variables may operate so that the influence of many variables can be separated from the variable of interest. Stated another way, an investigation helps rule out or to make implausible the influence of many variables that might explain changes on the dependent measures.

The extent to which an experiment rules out as explanations those factors that otherwise might account for the results is referred to as internal validity. Factors or sources of influence other than the independent variables are referred to as threats to internal validity and include history, maturation, testing, instrumentation, statistical regression, selection biases, attrition, combination with other threats (e.g., selection $\times$ history), diffusion of treatment, and special treatment or reactions of controls.

Aside from evaluating the internal validity of an experiment, it is important to understand the extent to which the findings can be generalized to populations, settings, measures, experimenters, and other circumstances than those used in the original investigation. The generality of the results is referred to as the external validity. Although the findings of an investigation could be limited to any particular condition or arrangement unique to the demonstration, a number of potential limitations on the generality of the results can be identified. These potential limitations are referred to as threats to external validity and include characteristics of the sample, the stimulus conditions or setting of the investigation, reactivity of experimental arrangements, multiple-treatment interference, novelty effects, reactivity of assessments, test sensitization, and timing of measurement.

Internal and external validity address central aspects of the logic of experimentation and scientific research more generally. The purpose of research is to structure the situation in such a way that inferences can be drawn about the effects of the variable of interest (internal validity) and to establish relations that extend beyond the highly specific circumstances in which the variable was examined (external validity). There often is a natural tension between meeting these objectives. Occasionally, the investigator arranges the experiment in ways to increase the likelihood of ruling out threats to internal validity. In the process, somewhat artificial circumstances may be introduced (e.g., videotapes to present the intervention, scripts that are memorized or read to the subjects; exposing nonhuman animals to interventions that vaguely reflect the phenomenon as it appears on the world). This means that the external validity may be threatened. The purposes of the investigation, both short and long term, ought to be clarified before judging the extent to which the balance of threats is appropriate. Sometimes external validity is not a major concern, especially in the context of testing theoretical predictions and determining whether something can occur under even rather special circumstances (proofs of concept). Indeed, some of the most intriguing research may be basic studies in a context that is unusually artificial so that critical processes can be revealed.

Internal and external validity are concepts to include in a methodological thinking tool kit. The threats are a way to evaluate any study.

Critical Thinking Questions

1. Why is priority usually given to the internal validity rather than external validity?
2. What is a proof of concept study, and why is such a type of study important? Give an example (real or hypothetical) of a proof of concept study or finding.
3. In criticizing a study, it is not quite fair to say, "Yes but the results may not generalize to this or that population." What more is needed to make this a genuine threat to external validity?

Chapter 2 Quiz: Internal and External Validity

Chapter 3
Construct and Data-Evaluation Validity

Learning Objectives

3.1 Define construct validity

3.2 Analyze the reasons that make construct validity intriguing

3.3 Examine the clinically identified threats to construct validity

3.4 Analyze basic threats as the first step to manage construct validity

3.5 Assess the utility of the statistical evaluation of construct validity

3.6 Review the threats to data-evaluation validity

3.7 Review some primary concepts of data-evaluation validity

3.8 Analyze some major threats to data-evaluation validity

3.9 Explain the importance of threats to data-evaluation validity in the planning stage

3.10 Identify ways to address the problems faced during experiments to obtain the best outcome

We have discussed internal and external validity, which are fundamental to research. Two other types of validity, referred to as construct validity and data-evaluation validity, also must be addressed to draw valid inferences.[1] All four types relate to the conclusions that can be reached about a study. Construct and data-evaluation validity are slightly more nuanced than are internal and external validity. They are more likely to be neglected in the design of research in part because they are not handled by commonly used practices. For example, random assignment of subjects to various experimental and control conditions nicely handles a handful of internal validity threats (e.g., history, maturation, testing, and selection biases), and we are all generally aware of this. Also, for external validity, we are all routinely concerned about and aware of the generality or lack of generality as a potential problem. Less in our awareness are the major threats that comprise this chapter.

Construct validity often relates to interpretation of the findings and whether the investigators can make the claims they wish based on how the study was designed. Data-evaluation validity takes into account subtle issues about analysis of the data. This is subtle too because we often teach data analysis as a mechanical procedure—i.e., OK, I collected my data, now what statistics do I use. Yet, data evaluation and statistical issues can undermine the study before the first subject is ever run! That is, the planning of the study can include procedures or practices that will make it very likely that no differences between groups will be evident even if there is really an effect. (In retrospect, I can see now that I should never have said to my dissertation committee that I did not want my study to be bogged down with construct or data evaluation conclusion validity or for that matter control groups.) This chapter considers construct and data-evaluation validity and the interrelations and priorities of the different types of validity. The goal is to describe the nature of these types of validity and the threats they raise.

3.1: Construct Validity Defined

3.1 Define construct validity

Construct validity has to do with interpreting the basis of the relation demonstrated in an investigation. A *construct* is the underlying concept that is considered to be the basis

for or reason that experimental manipulation had an effect. Construct validity might be ambiguous as a term, but as a guide here, think of this as interpretive validity. Problems of construct validity in relation to the present discussion pertain to ambiguities about interpreting the results of a study.

It is helpful to delineate construct from internal validity. Internal validity focuses on whether an intervention or experimental manipulation is responsible for change or whether other factors (e.g., history, maturation, testing) can plausibly account for the effect. Assume for a moment that these threats have been successfully ruled out by randomly assigning subjects to experimental and control groups, by assessing both groups in the same way and at the same time, and so on. We can presume now that the group differences are not likely to have resulted from the threats to internal validity but rather from the effects of the experimental manipulation. It is at this point that the discussion of construct validity can begin. What *is* the experimental manipulation or intervention, and *why* did it produce the effect?

Construct validity addresses the presumed cause or the explanation of the causal relation between the intervention or experimental manipulation and the outcome. Is the reason for the relation between the intervention and change due to the construct (explanation, interpretation) given by the investigator? For example, let us say that an intervention (e.g., psychotherapy for anxiety, pain medication after surgery) is better than no treatment.

> **CONSTRUCT VALIDITY ASKS:** Why did this difference occur? What was responsible for the superiority of the experimental group over some control group? Although it might seem obvious that it must have been the intervention, it turns out not to be obvious at all what facet of the intervention led to the change. That is why construct validity is more nuanced and likely to be neglected than stark in your face threats to internal validity.

3.2: Confounds and Other Intriguing Aspects of Construct Validity

3.2 Analyze the reasons that make construct validity intriguing

There are several features within the experiment that can interfere with the interpretation of the results. These are often referred to as *confounds*. When we say an experiment is confounded, that refers to the possibility that another variable co-varied (or changed along with or was embedded in the experimental manipulation) with the intervention. That confound could in whole or in part be responsible for the results. Some component other than the one of interest to the investigator might be embedded in the intervention and accounts for the findings.

For example, consider a familiar finding about the effects of moderate amounts of wine on health. Consuming a moderate amount of wine (e.g., 1–2 glasses with dinner) is associated with increased health benefits, including reduced risk of cardiovascular disease, some forms of cancer, Type 2 diabetes, and many other conditions (e.g., Guilford & Pezzuto, 2011; Opie & Lecour, 2007). In studies of this relation, consumption of wine is the construct or variable of interest. The basic test comes from a two-group comparison, namely, those who consume a moderate amount of wine and those who do not. Threats to internal validity are usually ruled out and just assume they are for the moment. Alas, the findings indicate consuming a moderate wine is associated with health benefits. Actually, it is useful to be more careful even in stating the finding: the group with moderate wine drinking had better health than the group that did not drink. The construct validity question usually begins with "yes, but." So here we ask, "*Yes* the groups are different, *but* is it the consumption of wine or something else?"

Maybe people who drink wine are generally mellower and easy going (even without the wine), more social, less likely to smoke cigarettes, and have lower rates of obesity than nonwine drinkers. Indeed, maybe while wine drinkers are sipping wine, their nonwine drinking controls are stuffing themselves with nacho chips and cheese and watching television. That is, wine drinking may be confounded (associated, correlated) with diet or some other variable(s), and these other variables, rather than the wine drinking, may account for the finding. It may be a package of characteristics, and the wine part may or may not contribute to better health. It makes a huge difference in knowing the answer. If it is not wine drinking, encouraging people to drink may have no impact on heart disease and health more generally. Indeed, we want to be especially careful because some who might take up moderate drinking will unwittingly escalate to heavier drinking, which can be quite harmful to health (e.g., increase the risk of heart disease and cancer). This question and concerns here pertain to construct validity. Taking into account these other variables that may explain the finding is more nuanced and may require selecting special control subjects and using statistical analyses to help isolate one particular influence (wine drinking).

In passing, the benefits of drinking moderate amounts of wine seem to hold, i.e., wine plays a role. But wine drinking is associated with (confounded by) other characteristics. People who drink wine, compared to those who drink beer and other alcohol (spirits), tend to live healthier life styles and to come from higher socioeconomic classes. They tend to smoke less, to have lower rates of obesity, and to be lighter drinkers (total alcohol consumption) (e.g., counting all beer and liquor). Each of these characteristics is related to health. Yet, even after controlling these, moderate wine drinking

contributes to health. So, have we resolved the construct validity questions? More can be asked to home in on the construct. What about the wine? Immediately, one turns attention to alcohol consumption, but studies have shown that the benefits can be obtained by removing the alcohol. Also, red and white wine have benefits but appear to work (at the biochemical level in the body) for different reasons (e.g., Siedlinski, Boer, Smit, Postma, & Boezen, 2012). All of this has been the subject of fascinating research, including fine-grained evaluation of the biochemical, molecular, and genetic level to understand how components operate on the body. Thus, one can continue to ask finer-grained questions about the "why and how" of an effect.

Construct validity is intriguing in part because it is at the interface of methodology (e.g., conducting well-controlled, careful studies) and substantive understanding (e.g., theory and evidence about what actually explains the phenomenon of interest). One can easily make a research career based on demonstrating the impact of some manipulation and then analyzing at various levels why that effect occurred, i.e., construct validity. The construct validity question is not about methodology alone, but about understanding the independent variable, and this is what theory, hypotheses, and predictions are about. In research, it is invariably useful for the investigator to ask, what might be going on here to cause the effects I am observing or predicting? Is there any way I can isolate the influence more precisely than gross comparisons of groups? The gross comparison of groups (e.g., wine drinkers vs. nondrinkers) is an excellent point of departure, but only a beginning in the effort to understand.

In one's own studies (and even more so in everyone else's studies!), it is meaningful to challenge the findings with the questions, what is the variable the investigator studied and could that variable include other components than those discussed by the investigator? Those features associated with the experimental manipulation that interfere with drawing inferences about the basis for the difference between groups are referred to as *threats to construct validity*.

3.3: Threats to Construct Validity

3.3 Examine the clinically identified threats to construct validity

Construct validity pertains to the reason why the independent variable has an effect. One cannot identify in the abstract or in advance of a study all the competing constructs that might be pertinent to explain a finding. Yet, there are a number of factors that emerge in many different areas of clinical research and can be identified as threats to construct validity. Table 3.1 summarizes major threats to construct validity to provide an easy reference.

Table 3.1: Major Threats to Construct Validity

Specific Threat	What It Includes
Attention and Contact Accorded the Client	The extent to which an increase of attention to the client/participant associated with the intervention could plausibly explain the effects attributed to the intervention.
Single Operations and Narrow Stimulus Sampling	Sometimes a single set of stimuli, investigator, or other facet of the study that the investigator considers irrelevant may contribute to the impact of the experimental manipulation. For example, one experimenter or therapist may administer all conditions; at the end of the study, one cannot separate the manipulation from the person who implemented it. In general, two or more stimuli or experimenters allow one to evaluate whether it was the manipulation across different stimulus conditions.
Experimenter Expectancies	Unintentional effects the experimenter may have that influence the subject's responses in the experiment. The expectancies of the person running subjects may influence tone of voice, facial expressions, delivery of instructions, or other variations in the procedures that differentially influence subjects in different conditions.
Demand Characteristics	Cues of the experimental situation that are ancillary to what is being studied but may provide information that exerts direct influence on the results. The cues are incidental but "pull," promote, or prompt behavior in the subjects that could be mistaken for the impact of the independent variable of interest.

3.3.1: Attention and Contact with the Clients

Attention and contact accorded the client in the experimental group or differential attention across experimental and control groups may be the basis for the group differences and threaten construct validity. This threat is salient in the context of intervention (e.g., psychotherapy, prevention, medicine) research. In these contexts, the intervention may have exerted its influence because of the attention provided, rather than because of special characteristics unique to the intervention.

A familiar example from psychiatric research is the effect of placebos in the administration of medication. Suppose investigators provide a drug for depression to some patients and no drug to other patients. Assume further that groups were formed through random assignment and that the threats to internal validity were all superbly addressed. At the end of the study, patients who had received the drug are greatly improved and significantly different from those patients who did not receive the drug. The investigator may then discuss the effect of the drug and how this particular medication affects critical biological processes that control symptoms of depression. We accept the *finding* that the intervention was responsible for the outcome; that is, groups were different and none of the internal validity threats are very plausible. Yet, on the basis of construct validity concerns, we may not accept the *conclusion*. (Finding is the

descriptive statement—groups were different; conclusion is the inference one draws or the explanatory statement.)

The intervention consists of all aspects associated with the administration of the medication in addition to the medication itself. We know that taking any drug might decrease depression because of expectancies for improvement on the part of the patients and on those administering the drug. Merely taking a drug and undergoing treatment, even if the drug is not really medication for the condition, can lead to improvement. This is called a *placebo effect*.

> *A placebo is a substance that has no active pharmacological properties that would be expected to produce change for the problem to which it is applied.*

A placebo might consist of a pill or capsule that is mostly sugar (rather than antidepressant medication) or an injection of saline solution (salt water) rather than some active medication for the problem of interest.

Placebo effects are "real," powerful, and important. They are studied in research and used clinically to make people better (Benedetti, 2009). For example, medications for depression are effective (lead to recovery, significantly reduce symptoms) in approximately 50–60% of the adult patients; placebos are effective for 30–35% of the patients (Agency for Health Care Policy and Research, 1999). Interestingly, in some studies talk therapy and medication do not surpass the effectiveness of placebos in treating depression (e.g., Barber, Barrett, Gallop, Rynn, & Rickels, 2012). The relative impact of placebo and medication for depression can vary as a function of severity of the disorder. For mild-to-moderate depression, medication and placebos are about equally effective in reducing symptoms, but when depression is severe, medication effects surpass the impact of placebos (Fournier et al., 2010).

> A "regular" placebo is a completely inactive drug with no real effects or side effects based on the chemical composition (e.g., sugar pill). Active placebos, on the other hand, consist of placebos that mimic some of the side effects of real medications.

The active placebo still has no properties to alter the condition (e.g., depression) but can produce side effects (e.g., dry mouth, blurred vision, nausea, sleep disruption) similar to the medication. When compared to active placebos, the differences between medication and placebos are relatively small (Moncrieff, Wessely, & Hardy, 2004). Active placebos are considered to be more effective because they increase expectations on the part of patients (but maybe others who run the study) that the treatment is real.

There are of course huge ethical issues in administering placebos, intentionally "fake" drugs. My comments here do not advocate fake treatments. Equally worth noting is the ethics on the other side of this issue; that is, "real" treatments can have side effects, be invasive, and be very costly, and may not be needed. For example, a study of surgery included patients with knee osteoarthritis, a progressive disease that initially affects the cartilage of the knee (Moseley et al., 2002). Patients were selected with at least moderate knee pain and who had not responded to treatment. Arthroscopy is the most commonly performed type of orthopedic surgery, and the knee is the most common joint on which it is performed. Does the surgery help? In this study, patients were randomly assigned to surgery to repair the joints or to "sham" surgery. The sham surgery did not focus on the knee. Rather, these patients received small skin incisions under a fast-acting tranquilizer. The results showed that surgery or sham patients did not differ at 1-year and 2-year postoperative assessments on subjective measures of pain, reports of general health, and objective measures of physical functioning (walking, stair climbing). Thus, special features of the usual operation were not required for improvement. Perhaps expectations on the part of patients who then experienced less pain and behaved differently could explain the effects.

Conclusions about placebos versus treatments can vary on different dependent measures. For example, in a study of the treatment of asthma patients received different conditions: inhaler with an active medication (albuterol), inhaler with placebo, sham (fake) acupuncture, or no intervention (Wechsler et al., 2011). On an objective measure of expiration (forced expiratory volume), the active treatment was better than the other conditions. However, on patient reports of improvement, there was no difference among the treatment, placebo, and sham conditions; all of those groups were no different from each other but better than no treatment. Presumably from a clinical standpoint, one would advocate the treatment that produced changes on objective measures and subjective measures rather than just the latter.

A more common pattern of results was evident in a study that randomly assigned with anxiety disorder to stress-reduction, mindfulness-based treatment or a wait-list control (no treatment) (Vøllestad, Sivertsen, & Nielsen, 2011). The wait-list group received no intervention. The results showed that the treatment group improved on measures of depression, anxiety, and worry compared with the control group. The control group nicely controls for threats to internal validity (e.g., history, maturation, testing, and regression). Yet, the study did not control for attention and expectations that are well known to be important in therapy studies. The authors conclude that the specific intervention was effective for anxiety disorders. In deference to construct validity, we say that the intervention was effective (finding) but that the conclusion that this specific intervention explained the change (conclusion) can be readily challenged. Attention and expectations in therapy studies are a quite plausible alternative hypothesis.

In general, expectancies for improvement can exert marked therapeutic effects on a variety of psychological

and medical dysfunctions. In relation to construct validity, expectancies for improvement must be controlled if an investigator wishes to draw conclusions about the specific effects of the intervention (e.g., medication, psychotherapy). In the case of medication, a placebo or active placebo permits control of expectancies. In psychological studies, a placebo condition is conceptually complex—what is a fake treatment? Procedures designed to have no real therapeutic effects when psychological treatments are compared can be as effective as "real" treatments, especially if those fake treatments generate expectations that the patient will improve (e.g., Baskin, Tierney, Minami, & Wampold, 2003; Boot, Stothart, & Stutts, 2013; Grissom, 1996).

Placebo effects can be prompted by those who administer the drug (physicians or nurses) and influence patient responses by their expectations and comments. Consequently, physicians and nurses as well as patients ought to be naive ("blind") to the conditions (medication or placebo) to which patients are assigned. (The word "masked" is sometimes used to replace "blind" to eschew any reference to visual impairment. Because "blind" remains pervasive in research methodology, the term is retained here.) A *double-blind study* is so called because both parties (e.g., staff, patients) are naïve with regard to who received the real drug.[2] The goal of using a placebo of course is to ensure that expectations for improvement might be constant between drug and placebo groups. With a placebo control group, attention and contact with the client and expectations on the part of experimenters of clients become less plausible constructs to explain the effects the investigator wishes to attribute to the medication. Yet, in a very large proportion of studies, doctors and nurses can readily identify who is in the medication group based on comments by the patients about side effects or some other experience. Hence, there is a concern that placebo and expectation effects are not invariably controlled even in placebo control studies.

In general, there is a threat to construct validity when attention, contact with the subjects, and their expectations might plausibly account for the findings and were not controlled or evaluated in the design. A design that does not control for these factors is not necessarily flawed. The intention of the investigator, the control procedures, and the specificity of the conclusions the investigator wishes to draw determine the extent to which construct validity threats can be raised. If the investigator presents the *findings* as the therapy group led to a better outcome than the control group, we cannot challenge that. If the investigator wishes to *conclude* why the intervention achieved its effects, attention and contact ought to be ruled out as rival interpretations of the results.

For example, a study was designed to reduce depression among individuals with multiple sclerosis who were carefully screened for depression (Bombardier et al., 2013). Adults were randomly assigned to receive a physical activity-based treatment or no treatment (wait list). Physical activity and exercise have been established as a viable treatment for depression, and some of the neurological underpinnings of how have been revealed as well (e.g., Carek, Laibstain, & Carek, 2011; Duman, Schlesinger, Russell, & Duman, 2008). In this study, treatment led to improved gains on some measures, but there were no differences on other measures. The authors discussed the effects of the treatment, but of course there is a construct validity problem.

> Did the specific treatment (physical activity) make a difference, or was it all of the attention, contract, and other accoutrements of treatment that made the differences? Expectancies and common factors of therapy exert fairly reliable impact, and not controlling these makes them a very plausible rival hypothesis.

In general, participation in any treatment can generate expectations for improvement, and these expectations serve as a parsimonious explanation of the results of many studies, especially when one group receives something and another group receives nothing (no treatment) or something paltry (very limited contact). Expectations are parsimonious because the construct provides an explanation of the effects of many studies in which treatment (e.g., some form of psychotherapy or medication) is better than a control condition.

Critical Thinking Question

Why have placebo effects and expectations been threats to construct validity in evaluating the effects of medication and psychotherapy, respectively?

3.3.2: Single Operations and Narrow Stimulus Sampling

Sometimes an experimental manipulation or intervention includes features that the investigator considers as irrelevant to the study. These features can include stimulus materials used in the experiment, a vignette or script presented on a laptop or tablet screen, or people (e.g., one research assistant) who run the subjects through the experimental procedures. All of these seem to be irrelevant but may influence the findings. The construct validity question is the same as we have discussed so far, namely, was the experimental manipulation (as conceived by the investigator) responsible for the results, or was it some seemingly irrelevant feature with which the intervention was associated?

For example, consider we are conducting an experiment and we use one experimenter who administers two conditions. We want to see if the conditions, variations of our experimental manipulation, differ. One experimenter provides both conditions and sees all of the subjects. At the end of the investigation, suppose that one condition is clearly different from the other on the dependent measures.

The investigator may wish to discuss how one condition is superior and explain on conceptual grounds why this might be expected. We accept the finding that one condition was more effective than the other. In deference to construct validity we ask, what *is* the experimental condition, i.e., what did it include? The comparison consisted of this one therapist giving both conditions. One might say that the experimenter was "held constant" because he or she was used in both groups. But it is possible that this particular experimenter was more credible, comfortable, competent, and effective with one of the conditions than with the other. Perhaps the experimenter had a hypothesis or knew the hypothesis of the investigator, performed one of the conditions with greater enthusiasm and fidelity than the other, or aroused patients' greater expectancies for change with one of the conditions. The differential effects of the two conditions could be due to the statistical interaction of the experimenter × group condition, rather than group condition along. Another way to say this was the experimenter combined with the manipulation was the "real" construct or at least we cannot rule it out. The study yields somewhat unambiguous results because the effect of the experimenter was not separable in the design or data analyses from the effects of the different groups.

The situation I have described here is evident in contemporary studies. Consider these examples.

Example 1

In one randomized controlled trial (RCT) for posttraumatic stress disorder and panic attacks, one therapist treated all of the patients (Hinton et al., 2005). It made sense for this in part because the patients were all Cambodian refugees and the therapist was fluent in the language. Yet from a methodological standpoint, we cannot separate the effects of treatment from effects of treatment as administered by this therapist. Perhaps she engaged in different behaviors apart from the treatment procedures or had different expectations between the two treatments. A second therapist would have made an enormous difference in the inferences that could be drawn. The authors concluded that one of the treatments was more effective than the other. We can state that the interventions were different (findings) but have to qualify the explanation (conclusion) because of a threat to construct validity.

Example 2

A laboratory experiment designed to evaluate opinions held about mental illness. The purpose is to see if people evaluate the personality, intelligence, and friendliness of others differently if they believe these other persons have been mentally ill. College students serve as subjects and are assigned randomly to one of two conditions. In the experimental condition, the students view a laptop screen where they see a 30-year-old man. They then listen to a prerecorded tape that describes him as holding a job at a software start-up company, and living at home with his wife and two children. The description also includes a passage noting that the man has been mentally ill, experienced strange delusions, and was hospitalized 2 years ago. In the control condition, students see the same man on the screen and hear the same description except the passages that talk about mental illness and hospitalization are omitted. At the end of the tape, subjects rate the personality, intelligence, and friendliness of the person they just viewed. Alas, the hypothesis is supported—subjects who heard the mental illness description showed greater rejection of the person than did subjects who heard the description without reference to mental illness.

The investigator wishes to conclude that the content of the description that focused on mental illness is the basis for the group differences. After all, this is the only feature that distinguished experimental and control groups. Yet, there is a potential construct validity problem here. The use of a single case on the screen (i.e., the 30-year-old man) is problematic. It is possible that rejection of the mental illness description occurred because of special characteristics of this particular case (e.g., sex, ethnicity, age, facial structure and expression). The difference could be due to the manipulation of the mental illness description *or* to the interaction of this description with characteristics of this case. We would want two or more persons shown (on the laptop) in the study who vary in age, sex, and other characteristics so that mental illness status could be separated from the specific characteristics of the case. Even two would be a great improvement in the methodology of this study. With two cases presented on the screen, the results could rule out (by statistical tests) that the case does not make a difference (assuming that it does not). That is, the effect of the experimental manipulation does not depend on unique characteristics of the case description. In general, it is important to represent the stimuli in ways so that potential irrelevancies (e.g., the case, unique features of the task) can be separated from the intervention or variable of interest. Without separating the irrelevancies, the conclusions of the study are limited.

Example 3

Finally, consider the case where two evidence-based psychotherapies are being compared. Let us say we recruit therapists who are experts in treatment A to administer that treatment and other therapists skilled in treatment B to administer that treatment. Thus, different therapists provide the different treatments. This is reasonable because we may wish to use experts who practice their special techniques. At the end of the study, assume that therapy A is better than B in the outcome achieved with

the patient sample. Because therapists were different for the two treatments, we cannot really separate the impact of therapists from treatment. We might say that treatment A was better than treatment B. Yet, perhaps therapists who administered treatment A may have simply been much better therapists than those who administered treatment B and that therapist competence may account for the results. The confound of treatment with therapists raises a significant ambiguity. Another way of saying this is to describe the independent variable. The investigator pitched the study to us as treatment A versus treatment B, but our OCMD colleague cogently points out that the study really examined treatment-A-as-practiced-by-therapist-team 1 versus treatment-B-as-practiced-by-therapist-team 2. As with any threat, the plausibility of considering this as a competing interpretation is critical. During the study, the investigator may collect data to show that somehow therapists were equally competent (e.g., in adhering to their respective treatments, in training and initial skill, and in warmth). Equivalence on a set of such variables can help make the threat to construct validity less plausible. Even so, some overall difference including therapist competence that is not assessed might, in a given study, continue to provide a plausible threat to construct validity.

The use of a narrow range of stimuli and the limitations that such use imposes sound similar to external validity. It is. Sampling a narrow range of stimuli as a threat can apply to both external and construct validity. If the investigator wishes to *generalize* to other stimulus conditions (e.g., other experimenters or types of cases in the above two examples, respectively), then the narrow range of stimulus conditions is a threat to *external validity*. To generalize across stimulus conditions of the experiment requires sampling across the range of these conditions, if it is plausible that the conditions may influence the results. If the investigator wishes to *explain why* a change occurred, then the problem is one of *construct validity* because the investigator cannot separate the construct of interest (e.g., treatment or types of description of treatment) from the conditions of its delivery (e.g., the therapist or case vignette).

The same problem in a study may serve as a threat to more than one type of validity. I have discussed narrow stimulus sampling as one example. Some problems (e.g., attrition) serve as threats to all types of validity. Also, the types of validity themselves are not all mutually exclusive. Construct and external validity, for example, can go together and which one to invoke in evaluating a study has to do with the particular conclusion one wishes to make or to challenge. In the previous example, if the investigator wishes to say that the mental-illness description led to the change, the skeptic can cogently say, "Maybe, but maybe not because of construct validity." If the investigator wishes to say that the findings apply to adults in general (e.g., women, the elderly, different ethnicities), the skeptic can cogently say, "Maybe, but maybe not because of external validity."

As I mentioned, any threat is really only a threat when it is a plausible account of the finding or can more parsimoniously account for that finding. Narrow stimulus sample may or may not be plausible in any of the instances I used to illustrate this particular threat. Even so, in a given study, it is useful to sample across a wider range of conditions presented in a given study. By "wider range" this can be only two or more experimenters who administer all of the conditions provided to the groups or two or more vignettes or sets of stimuli. This allows data analysis to see whether the manipulation was dependent on only one of the experimenters or stimuli material. The strategy is useful as well for external validity by showing that the results are not restricted to a very narrow set of conditions.

3.3.3: Experimenter Expectancies

In both laboratory and clinical research, it is possible that the expectancies, beliefs, and desires about the results on the part of the experimenter influence how the subjects perform.[3] The effects are sometimes referred to as *unintentional expectancy effects* to emphasize that the experimenter may not do anything on purpose to influence subjects' responses. Depending on the experimental situation and experimenter–subject contact, expectancies may lead to changes in tone of voice, posture, facial expressions, delivery of instructions, and adherence to the prescribed procedures and hence influence how participants respond. Expectancy effects are a threat to construct validity if they provide a plausible rival interpretation of the effects otherwise attributed to the experimental manipulation or intervention.

Previously, I mentioned placebo effects. When treatments (e.g., medication, surgery) and control (e.g., placebo, sham surgery) are compared, it is important to keep staff who collect the data (e.g., rate, measure, or evaluate improvement) or otherwise meet with patients directly (e.g., doctors, nurses, therapists who deliver the medication or assessment procedures) "blind" to the conditions to which patients are assigned. Thus, medications and placebos might be placed in identical packages that are coded, and only the investigator *not* in direct contact with staff who administer the treatments or with the patients knows the key that explains who was given what medication/placebo and when. Sometimes even the investigator does not know until the very end of the study when the codes (for who received what) are revealed. The reason is to eliminate the bias that might come from knowing the conditions

to which subjects are assigned. Expectancies on the part of those involved in the study might be communicated to the patients or clients and somehow influence the results.

Experimenter expectancies are similar but emerged in the context of laboratory experiments in psychology rather than intervention studies. The work was prominent decades ago, primarily in the context of social psychological research (Rosenthal, 1966, 1976). Several studies showed that inducing expectancies in experimenters who ran the subjects through various laboratory conditions could influence the results. That is, leading experimenters to believe how the results would come out somehow influenced subject performance.

> It is important to be aware of the prospect that individuals running a study might unwittingly influence the results if they have expectancies about the likely direction of effects. The notion of experimenter expectancies, as a threat to validity, is infrequently invoked for several reasons.

The first and foremost perhaps is that expectancies currently are not a plausible explanation in many laboratory studies. Procedures may be automated across all subjects and conditions, and hence there is consistency and fairly strong control of what is presented to the subject. Also, in many investigations, subjects participate through the Web (MTurk, Qualtrics) and have no direct contact with an experimenter in the usual sense.

Second, in many laboratory paradigms, expectancies are not likely candidates for influencing the specificity of a finding that is sought. For example, clinical neuroscience focuses on changes and activation changes in the brain following presentation of stimulus material (e.g., scenes of a significant other, emotional stimuli). Hypotheses relate to mechanisms of action and are not likely to be influenced by what the researcher is expecting on the way to the scanner. Are expectancy influences possible? Yes—the research assistant can always chat about the study on the way to the scanner, but not likely in the general case.

Third, how experimenter expectancies exert their influence is unclear. A likely explanation is that the experimenter differentially adheres to the procedures and introduces bias that way. This can be controlled by training of experimenters or by automating as much of the procedures as possible. Yet, now we know of priming (cues that are outside of the conscious awareness of a person) and the influence that these subtle cues in the environment can exert. It is possible in theory that subjects are definitely influenced by subtle cues not easily identified.

All of that noted, in a given situation, expectations on the part of the experimenter may plausibly serve as a source of ambiguity and threaten the construct validity of the experiment. Usually research assistants who run subjects, when that is the basis of the procedures, have a view of how the results should come out. As an investigator, it is useful to minimize these expectations by not providing explicit information about what one is expecting to find.

3.3.4: Cues of the Experimental Situation

Cues of the situation refer to those seemingly ancillary factors associated with the experimental manipulation and have been referred to as the *demand characteristics* of the experimental situation (Orne, 1962). Although the topic goes back decades, demand characteristics continue to be a topic of research and recognized as a source of bias (e.g., Allen & Smith, 2012; Damaser, Whitehouse, Orne, Orne, & Dinges, 2009). Demand characteristics include sources of influence such as information conveyed to prospective subjects prior to their arrival to the experiment (e.g., rumors about the experiment, information provided during subject recruitment), instructions, procedures, and any other features of the experiment. These other features may seem incidental, but they "pull," promote, or prompt behavior in the subjects. The change in the subjects could be due to demand characteristics rather than the experimental manipulation. That of course would be a construct validity problem.

The defining example to show the influence of cues in the experiment distinct from the independent variable focused on the role of demand characteristics in a sensory deprivation experiment (Orne & Scheibe, 1964). Sensory deprivation consists of minimizing as many sources of sensory stimulation as possible for the subject. Isolating individuals from visual, auditory, tactile, and other stimulation for prolonged periods has been associated with distorted perception, visual hallucinations, inability to concentrate, and disorientation. These reactions usually are attributable to the physical effects of being deprived of sensory stimulation. Yet, perhaps cues from the experimental situation might evoke or foster precisely those reactions mistakenly attributed to deprivation. This is an interesting hypothesis, but testing it requires separating real sensory deprivation from the cues associated with the deprivation.

An experiment was completed where subjects were exposed to the accouterments of the procedures of a sensory deprivation experiment but actually were not deprived of stimulation. Subjects received a physical examination, provided a short medical history, were assured that the procedures were safe, and were exposed to a tray of drugs and medical instruments conspicuously labeled "Emergency Tray." Of course, any of us would be alerted to all sorts of potential issues and problems that might arise in light of these seeming safeguards for our protection. Also, subjects were told to report any unusual visual imagery, fantasy, or feelings, difficulties in concentration, disorientation, or similar problems. They were informed that they were to be placed in a room where they could work in an arithmetic task. If they wanted to escape, they could do so by pressing a

red "Emergency Alarm." In short, subjects were given all sorts of cues to convey that strange experiences were in store.

The subjects were placed in the room with food, water, and materials for the task. No attempt was made to deprive subjects of sensory stimulation. They could move about, hear many different sounds, and work at a task. This arrangement departs from true sensory deprivation experiments in which the subjects typically rest, have their eyes and ears covered, and cease movement as much as possible. A control group in the study did not receive the cues preparing them for unusual experiences and were told they could leave the room by merely knocking on the window. At the end of the "isolation" period, the experimental group showed greater deterioration on a number of measures, including the report of symptoms characteristically revealed in sensory deprivation experiments. In this experiment, the cues of the situation when provided without any deprivation led to reactions characteristic of deprivation studies. By implication, the results suggest that in prior research deprivation experiences may have played little or no role in the findings.

A more current example reflects the problem in another context. Some research has suggested that chewing gum helps increase alertness and attention. In a recent study on the topic, either subjects were told that gum helped or hindered alertness or subjects were given no expectation (Allen & Smith, 2012).

Could demand characteristics explain these effects?

The current study investigated the effects of gum and demand characteristics on attention and reported mood over time. Participants completed measures of mood and attention, with and without chewing gum. To manipulate demand characteristics, they were told that the hypothesized effect of gum was either positive or negative, or not given a demand. Gum chewing increased attention to a task independently of demand characteristics (i.e., without regard to the induced expectations); demand also contributed based on the information provided about the supposed effect of gum. In short, the results found both an influence of gum chewing free from the demand characteristics and an effect of what subjects were told to expect. The results of this experiment are instructive because they show that all-or-none thinking about demand characteristics is risky. That is, we would not want to ask, is the result of an experiment due to the experimental manipulation or demand characteristics? This study shows that both contribute.

Consider a final example of an experimental paradigm and focus more common in clinical psychological research. This is an illustration of a hypothetical study designed to induce a sad or happy mood in subjects and to evaluate the impact of that mood on how individuals make attributions or what words they recognize on a task following the mood induction. At the end of the study, the results show differences between the two conditions. Is it possible that mood induction really led to the results? Could it be that all of the cues including the "message" of the mood induction led to performance? That is, somehow subjects recognized what was wanted and behaved consistently. Mood change was really not needed at all as a construct to explain the results. Perhaps, if we just told subjects, "pretend for just a moment you were sad, how would you respond to the following task?" Now give them the task, and the results might show that if they understand the purpose they will act in the same way even though they are not really sad and have had no special mood induction. Cues alone, in this case instructions, could lead to the same result.

Demand characteristics can threaten the construct validity if it is plausible that extraneous cues associated with the intervention could explain the findings. The above demonstration and example convey the potential impact of such cues. Whether these demand characteristics exert such impact in diverse areas of research is not clear. Also, in many areas of research, the independent variable may include cues that cannot be so easily separated from the portion of the manipulation that is considered to be crucial. For example, different variations of an independent variable (e.g., high, medium, and low) may necessarily require different cues. Perhaps different demand characteristics are inherent to or embedded in different variations of the experimental manipulation. In such cases, it may not be especially meaningful to note that demand characteristics accounted for the results.

When several conditions provided to one group differ from those provided to a control group, one might weigh the plausibility of demand characteristics as an influence. Perhaps an implicit demand conveyed to control subjects that they are not expected to improve from one test occasion to another. That is, demand may not operate only on experimental subjects, but also maybe in a direction of limiting changes that otherwise might occur in a control group. Presumably if cues were provided to convey the expectation of no change for the control group, experimental and control differences that are obtained might well be due to different demand characteristics between groups.

3.4: Managing Threats to Construct Validity

3.4 Analyze basic threats as the first step to manage construct validity

As with other threats to validity, the first step in managing construct validity is to consider the basic threats and whether they can emerge as the study is completed. This is not a perfunctory task because threats to construct validity may influence what control or comparison groups are

included in the study and hence of course affect the basic design. For example, a common practice in research in psychopathology is to identify a patient group (e.g., individuals who meet criteria for depression, or anxiety, or schizophrenia) and to compare them with "healthy controls," i.e., individuals with no clinical dysfunction. The goal is to understand how the target group and disorder (depression) reflect some key psychological or biological process (e.g., emotional regulation, reaction to a perceptual task; brain activation when given a task). It would be good to consider construct validity in advance of the study, and indeed another group might well be added. The construct validity problem is that at the end of the study, the investigator may wish to include that there is something special about depression because the depressed patients were different from the healthy controls. Yet, the study really cannot draw conclusions about depression—at least based on the design. It may be that individuals with any psychiatric diagnosis (or a bit of a stretch, any disability) would show the differences with healthy controls. If the investigator wishes to talk about a specific disorder, the proper control is needed (e.g., some other disorder) to show that the effect is indeed related to the disorder of interest. A key question underlying construct validity in particular, but relevant to other types of validity as well, is: what does the investigator wish to say when the study is completed? It is valuable to consider that before doing the study because what one wants to say can influence how the study will be done.

All threats to construct validity cannot be anticipated in a study because construct validity is about interpretation of the basis of the effects of some manipulation or intervention. These interpretations often draw on different theories, may require many studies, and are the bases for deeper understanding. Yet, we discussed several threats that can emerge, and some comments on how they can be managed.

Attention, expectations, and placebo effects were mentioned as threats. These are potentially potent influences in studies. If it is possible that one or more of these influences could explain the findings, it is critical to include a control condition. Easier said than done. We know that comparing your brand new innovative and wildly clever treatment for a disorder when compared to no treatment has a construct validity problem. It may not be your treatment at all but attention and expectations in the treatment group that were not part of the non–treatment group. Researchers who are aware of this often compare the new and improved treatment to usual clinical care. The idea being that each group (new treatment, old treatment) received something. Yet, in virtually all studies we have no idea whether the new treatment and the treatment as usual generated the same level of expectancies for improvement.

> How to address the matter? Either in pilot work or during the study, obtain some measure of the extent to which participants expect improvement once they learned about their treatment condition (e.g., after the first session). At the end of the study, one can see if expectations for improvement differ between the conditions and also correlate (statistically) expectations at the beginning of treatment with therapeutic change.

Experimenter expectancies are slightly different from the expectations for change generated in participants. In any experiment, we would like it so that the expectations of those running the subjects are not aware of the hypotheses of the study. As I noted, experimenter expectations may not exert influence in many situations either by the nature of the experimenter–subject interaction (e.g., there may be none) or by the dependent measures (e.g., blood glucose level to evaluate control of diabetes, carbon monoxide that is exhaled as a measure of recent cigarette smoking). Different ways of running subjects and different measures vary in their amenability to such influences.

If experimenter expectancies could influence the results, there are two ways of managing these. First, provide a standard expectation or statement to experimenters who run the subjects so that they at least hear a constant mindset from the investigator. This expectation is not about what the hypotheses are but might be a speech that conveys the importance of running the subjects correctly through conditions or how the findings will be important no matter how they come out. Some standardization of what experimenters are told may reduce variability among research assistants if there are two or more. Rather than let them fill in their hypotheses, perhaps standardize what they are told in running the subjects. Second, and as mentioned, with attention and expectations of subjects, one can measure through a questionnaire what the beliefs of the experimenters are and see if those expectations are different among experimenters and also relate (correlate) expectations with outcome to see if in fact they are related. Presumably if expectancies were inconstant with the results that would make less plausible expectancies operated.

Demand characteristics are like expectations in principle but refer to how the cues of a study might prime, dictate, or influence the results. That is, it is not the experimental manipulation but rather expectations that lead to reactions on the part of subjects. Again, the first task is merely to note whether this is relevant to the study. That is, at the end of the study, could demand characteristics explain the findings? If it is possible that the cues or the context of the study gives away the desired responses on the dependent measures, then this threat to construct ought to be controlled. Three ways of controlling or assessing the impact of demand characteristics are to see what effects these characteristics have on the dependent measures without giving or exposing individuals to the experimental manipulation. Table 3.2 summarizes three ways in which this is accomplished.

Table 3.2: Procedures for Evaluating Whether Demand Characteristics May Account for the Results

Procedure	What Is Done	How to Interpret
Post experimental Inquiry	Ask subjects at the end of an experiment about their perceptions about the purpose, what was expected, how they were "supposed" to perform.	If subjects identify responses that are consistent with expected performance (the hypothesized performance), this raises the possibility that demand characteristics may have contributed to the results.
Pre-inquiry	Subjects are exposed to the procedures (e.g., told what they are), see what subjects would do, hear the rationale and instructions, but not actually run through the study itself. They are then asked to respond to the measures.	If subjects respond to the measures consistent with predicted or hypothesized performance, this raises the possibility that demand characteristics could contribute to the results.
Simulators	Subjects are asked to act as if they have received the procedures and then to deceive assessors (naïve experimenters) who do not know whether they have been exposed to the actual procedures. Similar to Pre-inquiry except that subjects actually go through that part of the experiment, if there is one, in which experimenters or assessors evaluate subject performance.	If simulators can deceive a naïve experimenter, i.e., make them believe they have actually been exposed to the experimental procedures, this is consistent with the possibility that demand characteristics could contribute to the results.

Each method of evaluating demand characteristics assesses whether the cues of the experimental situation alone would lead to performance in the direction associated with the independent variable. If the cues of the situation do not lead subjects to perform in the way that they would when exposed to the experimental manipulation, this suggests that demand characteristics are not likely to account for the results.

The post experimental inquiry focuses on asking subjects about the purposes of the experiment and the performance that is expected of them. Presumably, if subjects are aware of the purpose of the experiment and the performance expected of them, they can more readily comply with the demands of performance. Hence, their responses may be more a function of the information about the experiment than the manipulation itself. With this method, subjects actually go through the real experiment and are asked questions afterward. It is possible that subjects may not have perceived the demand characteristics consciously but still have responded to them in the experiment. The cues of the experiment that dictate performance may be subtle and depend upon behaviors of the experimenter or seemingly irrelevant procedures.

With the pre-inquiry, subjects are not actually run through the procedures in the usual way. Rather, they are asked to imagine themselves in the situation to which subjects would be exposed. These subjects may see the equipment that will be used, hear the rationale or instructions that will be provided, and receive all of the information that will be presented to the subject without actually going through the procedures. Essentially, the procedures are explained but not administered. After exposing the subject to the explanations of the procedures and the materials to be used in an experiment, the subjects are asked to complete the assessment devices as if they actually had been exposed to the intervention. The task is to respond as subjects would who experienced the procedures. Pre-inquiry research can inform the investigator in advance of conducting further investigations whether demand characteristics operate in the direction of expected results derived from actually running the subjects. Pre-inquiry data also may be useful when compared with data from actually conducting the investigation and running subjects through the procedures. If the Pre-inquiry data and experimental data are dissimilar, this suggests that the cues of the experimental situation alone are not likely to explain the findings obtained from actually being exposed to the experimental condition.

> The use of simulators also can evaluate demand characteristics. Simulators are subjects who are asked to act as if they received the experimental condition or intervention even though they actually do not.

These simulators are then run through the assessment procedures of the investigation by an experimenter who is "blind" as to who is a simulator and who is a real subject (i.e., a subject run through the procedures). Simulators are instructed to guess what real subjects might do who are exposed to the intervention and then to deceive a "blind" experimenter. If simulators can act as real subjects on the assessment devices, this means that demand characteristics could account for the results.

If data from post-inquiry, pre-inquiry, or simulators and from "real" subjects who completed the experiment are similar, the data are consistent with a demand-characteristics interpretation. The consistency does *not* mean that demand characteristics account for the results. Both demand characteristics and the actual effects of the independent variable may operate in the same direction. The consistency raises issues for construct validity and interpretation of the basis for the findings. If the data from evaluation of demand characteristics and real subjects do not correspond, this suggests that the cues of the situation do not lead to the same kinds of effects as actually running the subjects.

Efforts to evaluate the role of demand characteristics are to be actively encouraged if demand is a plausible and conceptually interesting or important threat to construct validity. If demand characteristics generate results different

from those generated by subjects who completed the experimental conditions, interpretation of the findings can be clarified. If demand characteristics can threaten construct validity, it is useful to design experiments so that merely exposing subjects to the cues (irrelevancies) of the experiment is not plausible as an explanation of the results. This can be accomplished by controlling or holding fairly constant all of the cues or by designing experiments so that the predicted results are counterintuitive, i.e., go in a direction opposite from what experimental demands would suggest.

In terms of managing other threats to construct validity, I have not mentioned stimulus sampling. This is the case where very narrow sampling of stimulus conditions presented to the subjects or included in the experiment might introduce ambiguity in interpreting the findings. This is a potential threat to external validity (do the results generalize beyond the narrow conditions in which they were presented?) and construct validity (the experimental manipulation cannot be separated from the narrow or single stimulus conditions, and the combination of stimulus condition and manipulation may explain the result). The procedures to address this were discussed in relation to external validity and are the same here, namely, try to vary the stimulus conditions used to present the experimental manipulation if those are case material, vignettes, brief movies, or stimulus material that might have two rather than one version. Similarly, one assistant running the study might be supplemented by at least one more. At the end of the study, one can analyze whether the different stimulus materials or research assistants varied in their effects and separate irrelevant parts of the experiment (e.g., how the manipulation was presented and who presented it) from the manipulation (what the key construct is underlying the experiment).

3.4.1: General Comments

The discussion has noted common threats to construct validity. However, a complete list of construct validity threats cannot be provided. The reason is that the threats have to do with interpretation of the basis for the results of an experiment. Thus, theoretical views and substantive knowledge about how the experimental manipulation works or the mechanisms responsible for change are also at issue, apart from the issue of experimental confounds. The questions of construct validity are twofold.

1. What *is* the independent variable (experimental manipulation, intervention)?

This question emphasizes the fact that the independent variable may be confounded with or embedded in other conditions that influence and account for the findings.

2. Why did that lead to change?

This question emphasizes the related issue of interpretation of what led the performance on the dependent measures. Here we do not speak of confound as much as better understanding of the mechanism, process, or theory to explain the change.

The questions encompass construct validity because they affect interpretation of the basis of a given finding.

Much of psychological research is devoted to understanding how a particular construct operates. The construct of interest is the focus with an effort to control artifacts or other constructs (e.g., expectancies of the experimenters, cues of the situation that are not of interest) that can obscure interpretation. What is the main construct and what is an artifact to be controlled are a matter of the investigator's interest. For example, if I am interested in studying the effects of medication, then patient expectancies are something I wish to control (by having placebo controls in the study). If I am interested in patient expectancies, then I want to control any extraneous medications they are on. The point is to convey that construct validity reflects an interplay and overlap of methodology (control of some variables) and substantive issues (so one can study another variable of interest).

3.5: Data-Evaluation Validity Defined

3.5 Assess the utility of the statistical evaluation of construct validity

Internal, external, and construct validity and their threats codify many of the concerns to which methodology is directed. The list of these concerns is long, so what more can remain? Actually a great deal. Assume we have designed our wonderful experiment to address the bulk of those threats already highlighted. Will the evaluation of our data reveal there is an effect or differences between groups? Whether we find differences between experimental and control conditions depends in part on whether there really are differences in the world between those conditions. Yet, even if there really are differences, whether we find those in our experiment depends on multiple considerations.

> *Data-evaluation validity* refers to those facets of the evaluation that influence the conclusions we reach about the experimental condition and its effect.[4]

In the vast majority of studies in the social, biological, and natural sciences, statistical analyses are used to evaluate the data and serve as the basis of drawing conclusions about whether an effect was evident. This is why data-evaluation validity has been previously referred to as statistical conclusion validity. Yet, the broader term "data evaluation" is of use because more can lead us astray than the maze of statistical analyses.

3.6: Threats to Data-Evaluation Validity Defined

3.6 Review the threats to data-evaluation validity

Statistical evaluation often is taught from two standpoints.

1. The first of these pertains to understanding the tests themselves and their bases. This facet emphasizes what the tests accomplish and the formulae and derivations of the tests (e.g., probability theory, distributions).
2. The second and complementary facet pertains to the computational aspects of statistical tests. Here concrete application of the tests to data sets, use of software, and interpretation of the findings are emphasized.

There is a third facet that might be considered at a higher level of abstraction, namely, the role of statistical evaluation in relation to research design and threats to validity. Data-evaluation validity reflects this level of concern with that evaluation and often is the Achilles' heel of research. This type of validity is often neglected when studies are planned and executed. That is, many (but not you or me of course) think of data analysis as something to consider once all the subjects are run and the numbers are in. So much is too late by then! There are several facets of the results and statistical evaluation that can obscure interpretation of the experiment. These are referred to as *threats to data-evaluation validity*.

It is important to note at this point that the discussion makes critical assumptions that are not fully agreed on in science. The assumptions are that statistical tests and probability levels (alpha), at least as currently practiced, are a good and reasonable basis for drawing inferences. These assumptions are a matter of debate (see Schmidt, 2010; Stang, Poole, & Kuss, 2010). In the present discussion, these issues are skirted in recognition of the fact that the bulk of research in psychology is based on drawing inferences from statistical evaluation. As such, there are common weaknesses of research that can be identified under the rubric of data-evaluation validity.

3.7: Overview of Essential Concepts of Data-Evaluation Validity

3.7 Review some primary concepts of data-evaluation validity

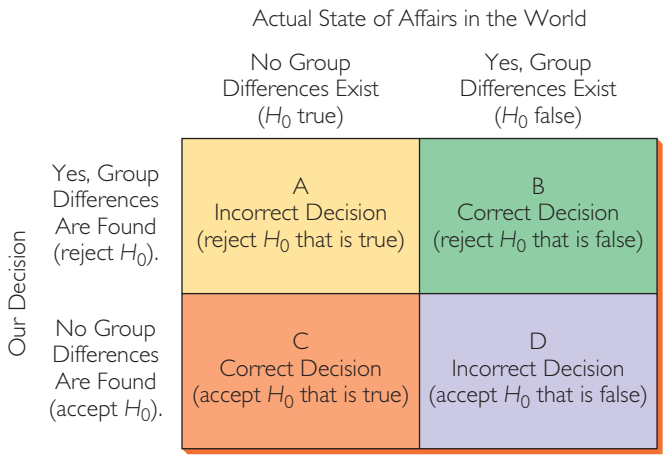
Before discussing the threats to validity, it is important to review a few of the essential concepts of statistical evaluation. As the reader well knows, in most psychological research, the conclusions in an experiment depend heavily on hypothesis testing and statistical evaluation.

> *The null hypothesis specifies that there are "no differences" between groups (e.g., experimental vs. control group).*

3.7.1: Statistical Test and Decision Making

Statistical tests are completed to evaluate whether the differences that are obtained are reliable or beyond what one is likely to find due to chance fluctuations. We can reject the null hypothesis of no difference if we find a statistically significant difference or accept this hypothesis if we do not. The rejection and acceptance of hypotheses are weighty topics only part of which we can treat here. The decision-making process is based on selecting a probability level that specifies the degree of risk of reaching a false conclusion. If the statistical difference between groups surpasses this probability level, we state that the difference is reliable and represents an effect of the experimental manipulation. If the difference fails to pass the threshold, we say that the difference is not statistically significant and that the groups are not different. Figure 3.1 notes the outcomes of an investigation based on the conclusions we might draw from statistical evaluation.

Figure 3.1: Actual State of Affairs in the World



The four cells represent the combination of *our decision* (we decide there is a difference vs. there is no difference) and the *true state of affairs in the world* (whether there really is a difference or there is no difference). Our goal in doing a study is to draw conclusions that reflect the true state of affairs in the world. That is, if there is a difference (e.g., in means) between two or more conditions (i.e., if the experimental manipulation made a difference), we

wish to reflect that in our decision (Cell B). If there is no difference between the conditions in the world, we would like to conclude that as well (Cell C). Occasionally, there is a clear (statistically significant) effect in our study, when in fact there really is no effect in the world (Cell A) or no effect in our study when in fact there is one in the world (Cell D). We specify our probability level (alpha) as the criterion for our decision making, i.e., concluding the difference we obtain is significant. By doing so, we also fix the risk of concluding erroneously that there is a difference when in fact there is none in the world and of concluding that there is no difference when in fact there is. The cells in Figure 3.1 have well-established names that reflect critically important statistical concepts to refer to the decision-making process, outcomes of our experiment, and risk of reaching a false conclusion. (The terms also make for terrific exam questions.) Table 3.3 lists these and other concepts that we draw on later to elaborate the threats to data-evaluation validity and to discuss of statistical evaluation more generally.

3.7.2: Effect Size

Among the concepts listed in Table 3.3, effect size is especially critical because it underlies several issues we shall consider.

Table 3.3: Important Concepts That Underlie Statistical Tests and Data-Evaluation Validity

Concept	Definition
Alpha (α)	The probability of rejecting a hypothesis (the null hypothesis) when that hypothesis is true. This is also referred to as a Type 1 error (Cell A).
Beta (β)	The probability of accepting a hypothesis (the null hypothesis) when it is false. This is also referred to as a Type II error (Cell D).
Power	The probability of rejecting the null hypothesis when it is false or the likelihood of finding differences between conditions when, in fact, the conditions are truly different. This probability is 1 – β (Cell B).
Effect size	A way of expressing the difference between conditions (e.g., treatment vs. control) in terms of a common metric across measures and across studies. The method is based on obtaining the difference between the means of interest on a particular measure and dividing this by the common (pooled) standard deviation.
Standard deviation	A measure of variation or variability about a mean. The standard deviation (also the square root of the variance) of a sample is given by the formula: $S = \sqrt{\frac{\sum (X_{i-n} - \overline{X})^2}{N - 1}}$ or $\frac{SS}{df}$ where X_{i-n} = individual observations of subjects *i* through *n* (all subjects); $\overline{X}$ = mean of the sample; N = sample size; SS = sum of squared deviation; df = degree of freedom

Effect size (ES) *refers to the magnitude of the difference between two (or more) conditions or groups and is expressed in standard deviation units.*

For the case in which there are two groups in the study, ES equals the differences between means, divided by the standard deviation:

Pooled standard variation is based on both groups combined as if they were one group and obtaining the standard deviation from that larger group.[5]

ES is expressed by the following equation.

$$ES = \frac{m_1 - m_2}{S}$$

For example, in a two-group study that evaluates treatment for clients experiencing anxiety, assume clients are assigned to treatment or no-treatment conditions. After the study, clients complete a measure of anxiety in which higher scores equal higher levels of anxiety. Suppose that treated subjects show a post treatment mean of 10 on the scale, whereas control subjects show a score of 16. We shall also suppose that the standard deviation is 8. ES equals .75 (derived from 10 minus 16 divided by 8). This means that in standard deviation units, the mean of the treatment group was .75 higher than the mean of the control group.

When ES is first taught and learned, emphasis is accorded what it means (magnitude or strength of effect) and how to compute it. From a methodological perspective, ES has a much broader role in a study. ES is equivalent to a bucket where many methodological problems and shortcomings of a study collect like dirty oil spilling from the bottom of a car. The more methodological problems, the smaller the ES and the less likelihood of showing statistically significant effects. In other words, whether one cares about the actual statistic of ES, in fact the methodological issues behind the statistic are of concern to anyone who does research. A little more detail is needed.

ES often is assumed to reflect the magnitude of the difference, as that difference exists in nature. Thus, if an investigator is exploring a truly potent variable, this will produce a marked ES and statistically significant results. However, ES is very much dependent on the design and methodology of the study in addition to a "true" state of affairs. A poorly planned or executed study can produce small and non–detectable effects even when the ES in nature is rather large. Not only flagrant methodological flaws, sloppiness, and error within the experiment but also more subtle nuances related to the procedures, subjects, and conditions can increase variation (the standard deviation) and dilute, diminish, and negate any differences that might otherwise be evident between groups. We will be talking about ES, but the impact on sloppiness that influences ES directly influences statistical significance and data-evaluation validity as well.

As investigators, we can influence ES in two general ways. First, if one looks at the ES formula, the numerator includes the difference between means of the groups included in the study. So one way to influence ES in a study is to be very thoughtful about what groups are included. As a general rule, select different levels of the variable of interest that are most likely to make the means quite different (e.g., very high vs. very low) in relation to your hypotheses.

> It is not only fine to select conditions that will maximize the likelihood of showing effects (large mean differences) but also prudent to do so. With a very strong test, positive or negative results may then be more likely to be interpretable.

For example, if we hypothesize that cholesterol is related to heart disease, we could compare two groups, individuals with "normal" levels versus individuals with slightly elevated levels of cholesterol and examine the proportion of individuals who have heart disease. The ES is likely to be lower and a statistically significant difference (in heart disease) is more difficult to demonstrate than if the study compared "normal" (or even low) levels and very elevated levels of cholesterol. Also, with logic that is often but not invariably correct, if really high levels of cholesterol produce no effect, then it is less likely that low levels will. (The logic depends in part on a linear relation between the variables, in this case cholesterol and heart disease. A linear relation happens to characterize the relation of cholesterol and heart disease, but such relations are not always the case.) In any case, the first way to increase ES and also the likelihood of obtaining a statistically significant result is to use conditions (groups) that are as discrepant as possible in likely outcomes within the constraints of your hypotheses. That is, we want to spread out the means and increase the predicted mean differences (numerator of the ES formula).

Second, ES can be greatly influenced and controlled by attending to the denominator of the ES formula, namely, the measure of variability. As a general statement, we can greatly influence the ES obtained in our study by reducing variability in the procedures to minimize the error term (standard deviation) that is the denominator in the ES equation. Many efforts to control features of the experiment are designed to minimize "error" variance, i.e., variability, in the formula for ES. The larger the variability (denominator), the smaller the ES for a constant difference between means (numerator). Later we shall talk about ways to reduce the denominator and increase the strength of the experimental test. For the moment, now armed with a few critical issues underlying statistical evaluation, we can talk about problems that interfere with data-based problems in drawing valid conclusions.

3.8: Threats to Data-Evaluation Validity

3.8 Analyze some major threats to data-evaluation validity

Several features of a study can undermine data-evaluation validity. They are slightly tricky because, as I mentioned, data evaluation is taught to be something one does once the data are collected. That is one first runs the study and then with all those numbers from all those measures, one meets with one's advisor and asks, "How do I analyze the data?" The tricky part is that data evaluation issues and threats to validity begin before the first subject is even run and before any number in the study is collected as a data point. In fact, by the time the first subject is run some of the problems (threats to validity) are already in place but just lurking like methodological bed bugs waiting in silence until the investigator climbs into bed and analyzes the data. Table 3.4 summarizes major threats to data-evaluation validity for easy reference but each is discussed here.

3.8.1: Low Statistical Power

Central to statistical evaluation is the notion of statistical power, which refers to the extent to which an investigation can detect differences between groups when differences exist within the population (see Table 3.3).

> *Power is the probability of rejecting the null hypothesis (i.e., there are no differences) when that hypothesis is false.*

Stated differently, power is the likelihood of finding differences between conditions when, in fact, the conditions are truly different in their effects. Certainly, if there is a difference between groups and if the experimental manipulation is effective, we wish to detect this difference in our statistical tests.

The central threat and probably most common threat to data-evaluation validity is relatively weak power or a low probability of detecting a difference if one truly exists. When power is weak, the likelihood that the investigator will conclude there are no differences between groups is increased. There might well be no differences in the world, and the intervention may in fact be no different in the effects it produces from those of a control condition. However, the conclusion of "no difference" might be due to low power, rather than to the absence of a difference between groups. The study must be designed so as to detect a difference if there is one.

Power is not an esoteric concept of relevance only to researchers in the confines of their studies, but can also affect decision making about practices that affect our daily lives and actually is a matter of life and death. For example, studies of whether screening makes a difference

Table 3.4: Major Threats to Data-Evaluation Validity

Specific Threat	What It Includes
Low Statistical Power	Power is the likelihood of demonstrating an effect or group difference when in fact there is a true effect in the world. Often studies have power that is too low to detect an experimental effect. Thus, no-difference finding could be due to the lack of a true effect or a study with too little power.
Subject Heterogeneity	Subjects recruited for a project will vary naturally in many ways. Yet, the extent of that variability can influence the conclusions that are drawn. If subjects can vary widely (in age, ethnicity, diagnoses, background, and so on), the variability (denominator in the effect size formula) also increases. As that variability increases, a given difference between groups (numerator in the effect size formula) becomes more difficult to detect. Generally it is advisable to specify the subject characteristics of interest and note inclusion and exclusion criteria so that variation is not unlimited.
Variability in the Procedures	How the study is executed can make a difference in whether a true effect is detected. If the procedures (e.g., in running a subject) are sloppy or inconsistent from subject to subject, unnecessary and undesirable variability is increased. And as with other threats related to variability that can interfere with detecting a difference when there is one.
Unreliability of the Measures	Error in the measurement procedures that introduces variability can obscure the results of a study. Measures that are not very reliable increase error in the assessment and as other sources of variability decrease the likelihood of showing group differences.
Restricted Range of the Measures	A measure may have a very limited range (total score from high to low) and that may interfere with showing group differences. The scores cannot spread out all of the subjects because of the limited range. No differences in a finding might be the result of the restricted range of the measure that could not permit a large enough scale to differentiate groups.
Errors in Data Recording, Analysis, and Reporting	Inaccuracies in data recording, analysis, and reporting refer to multiple steps in which inaccuracies enter into the database or the data are used in a selective way where only some measures or analyses are reported. Errors and selective reporting obviously mislead, whether intentional or unintentional, and threaten the data-evaluation validity of the study.
Multiple Comparisons and Error Rates	When multiple statistical tests are completed within the same investigation, the likelihood of a "chance" finding is increased. This is a threat to data evaluation because false conclusions will be more likely unless some accommodation is made for the number of tests (e.g., by adjusting the *p* level across the many tests to take into account the number of tests).
Misreading or Misinterpreting the Data Analysis	The conclusions reached from the data analysis are not to which the investigator is entitled. Either the proper statistic was not run or the conclusion reached goes beyond the statistical test.

for detecting cancer and the impact of cancer treatments on mortality occasionally have been unable to demonstrate differences due to weak statistical power (see Kramer, Berg, Aberle, & Prorok, 2011; Schutz, Je, Richards, & Choueiri, 2012). Similarly, in medication trials for a variety of diseases and conditions, low statistical power has been identified as a likely or possible reason of no differences (e.g., Tsang, Colley, & Lynd, 2009). The conclusion of low statistical power as a threat to validity can be identified in broad areas beyond cancer and other diseases (e.g., psychology, education, dentistry, and more). I have elected medical instances to convey more starkly that low power affects life and death decisions. While we are waiting for effective treatments for life-threatening diseases, we do not want to hear that viable treatments might work, but one could not tell because power was low within the studies!

More central to clinical psychology, comparisons of different psychotherapy techniques often show no differences in treatment outcome. Studies comparing two or more treatments have way too little power to detect "real" differences, given the relatively small samples and small ESs that characterize this research (Kazantzis, 2000; Kazdin & Bass, 1989). The problem is evident when treatment is compared to placebo controls too. For example, is medication more effective than placebos in treating depression among elderly patients? Could be, but a review of the available clinical trials concluded, "All of the trials were significantly underpowered to detect differences, resulting in inconclusive findings" (Nelson & Devanand, 2011, p. 577). Unfortunately, weak power characterizes studies in many areas of research within psychology (Maxwell, 2004).

Weak power has broad implications insofar as it slows theoretical and empirical advances (by misguiding us in the conclusions that are reached) and utilizing resources (subject and investigator time, tax dollars from grants) that might be more wisely used elsewhere. There are ethical implications as well: Is it ethical to subject participants to any procedures as part of an investigation if that investigation has very little likelihood of detecting a difference, even if there is one? Understandably, many funding agents require that grant applications include estimates of power to at least ensure that we as investigators think about the matter as we design, and long before we run, the study.

Whenever no difference (statistical) is evident between or among conditions, the first question to ask is whether the study had sufficient power to detect a difference if one were evident. Stated more colloquially, when you say to your advisor "this difference did not come out," the first statement back to you should not be "welcome to the group" but rather "are you sure you had sufficient power to detect a difference if there was one?"(And you should say back to your advisor, "where were you on this issue when I was planning the study?") Power is a critical issue, topic, and matter to resolve before a study is conducted.

3.8.2: Subject Heterogeneity

The notion of ES is useful as a way of introducing other threats to data-evaluation validity. Consider as a hypothetical experiment, a comparison of two groups or conditions (A and B), which are administered to different groups. Ordinarily, ES is considered to be a function of the true differences in the effects of these experimental and control conditions. That is, if condition A is more effective than condition B in the "real world," this will be evident in our experiment and be shown in our statistical evaluation. As I mentioned, the denominator of the ES formula includes a measure of variability (standard deviation). Thus, whatever outcome difference (on the dependent measures) between conditions A and B in our study, that difference will be influenced by variability in our experiment. This variability includes individual differences among the subjects (e.g., in personality, age, and IQ), random fluctuations in performance on the measures (e.g., errors subjects make, response styles in completing the measure, and mood and feelings on that day), differences in experimenters (i.e., research assistants) in how they administer the conditions, and other sources, not all of which are easily specifiable.

Standard deviation is a familiar statistical concept in relation to describing a sample and conducting statistical tests. This discussion is quite related to those tests but is more about how a study is carried out and the influence of that on the standard deviation and the likelihood of obtaining statistical significance if there is a real effect. (I focus on standard deviation here, but comments apply as well of course to the variance, which is the standard deviation squared.) The standard deviation is the "home" for all sorts of influences related to how a study is carried out.

With this overview, let us consider heterogeneity of the subjects and how what seems like a good thing can threaten validity. Subjects in an investigation can vary along multiple dimensions and characteristics, such as sex, age, background, race and ethnicity, and marital status. In the general case, the greater the heterogeneity or diversity of subject characteristics, the less likelihood of detecting a difference between conditions. Critical to the statement is the assumption that subjects are heterogeneous on a characteristic that is related to (or correlated with) the effects of independent variable. For example, clients who are recruited for a cognitive behavior therapy study may vary widely (and "lengthily") in shoe size. Is this heterogeneity of great concern? Probably not. It is unlikely treatment effects will relate to shoe size.[6]

As for other more directly pertinent variables, clients may vary widely in their severity or duration of the clinical problem, socioeconomic class (which correlates with general physical and psychological health), and other problems (e.g., substance abuse, depression, chronic medical disease) not of interest in the study. The impact of treatment and performance on the dependent measures might well be influenced by these factors. That these factors influence outcome is not inherently problematic or undesirable. However, heterogeneity of the sample means that there will be greater variability in the subjects' reactions to the measures and to the intervention. This variability will be reflected in the denominator for evaluating ES. As mentioned before, the greater that variability (denominator), the lower the ES for a given difference between means (numerator) and the less likely the difference in means will be statistically significant.

Consider as an example, a study in which individuals are recruited because they are depressed. We are going to compare those participants with others who are not depressed. Consider only our selection criteria for the depressed subjects. Screening criteria are invoked to ensure that subjects meet criteria for a psychiatric diagnosis of major depression. Consider three extraneous sources of variation that might be valuable to control. These are extraneous because they are not of interest in the study but can add to variability that will make it more difficult to detect group differences.

1. First, some of these subjects may also meet criteria for other psychiatric disorders as well (e.g., anxiety disorder, antisocial personality disorder). This is relatively common in part because meeting criteria for one disorder increases the likelihood of meeting criteria for another. Also, diagnoses are not that clean and distinct for many disorders. (Co-morbidity is the term used to refer to instances when an individual meets criteria for two or more disorders.) So one course of variability is the "other conditions" subjects bring.
2. Second, some of the depressed patients may be on medication. (The word "some" gives chills to methodologists because it means there is variability that might have controlled.) Among the participants on medications, the specific medication (there are many), how well those medications are monitored by a professional to ensure the dose is correct, and how well patients follow or adhere to taking their meds will vary as well. The diversity of these other treatments leads to increased variability (larger standard deviation).
3. Third, we left out something really obvious—age. The study might include anyone who is depressed and an adult (e.g., let us say 18 to 65). Such a wide age range—do we want that? Age differences this large can make for a highly variable sample because of the many other variables associated with age (e.g., psychological, biological). We do not need to specify them to note that they are related to variability. In fact, statistically one measure of variability is called the *range*, *which of course*

is the highest minus the lowest value on a given measure. Obviously 18–65 is a large age range, which means more variability.

We could go on with other sources of variability. For example, consider subject sex:

- Do we want males and females?
- What about gender identity?
- Do we include all?
- What about the type of depression?
- Do we want all people who meet some diagnostic cut-off for depressive symptoms (e.g., unipolar, postpartum)?
- What if individuals are depressed for different reasons (e.g., disability is a much greater source of depression in the elderly, some of whom are in this study)? And so on.

For all of the sources of variation I mentioned and for these latter questions, there is no single and certainly no correct answer. As one designs a study, the questions are: what provides the best (optimal) test of my hypotheses and what sources of variation can I control to provide the most sensitive test that can detect differences if there are any?

Critical Thinking Question

Why is variability (in procedures, subjects, measures) a threat to data-evaluation validity?

3.8.3: Variability in the Procedures

Variability in the procedures operates in the same way as a threat to data-evaluation validity as did subject heterogeneity. The effect is to increase the standard deviation in the ES formula and possibly dilute, weaken, and make significant differences between means more difficult to detect. It is more subtle only because we think less about this type of variability in comparison to variability in selecting subjects.

Ideally, the procedures will be held relatively constant and so as to minimize variation in how all facets of the study are implemented among subjects. Training experimenters to administer the instructions, standardizing the materials presented to subjects, and standardizing the collection of assessment data can minimize extraneous variation. Currently many studies are conducted online where stimulus materials are presented automatically. In addition, in many studies assessments are completed online through survey methods or software in which all measures are placed online and answered by subjects from a computer. The standardization in this way is useful, even though other problems related to variability can arise. For example, if subjects complete measures in their own homes on computer, what else is going on in the home (dog barking, children require care, and blaring rap music in the background) that might vary among subjects and provide less than consistent testing conditions. There is only so much one can control. Yet variation in procedures as a threat to validity encourages us as investigators to try. Rigor in the execution of the procedures is not a methodological nicety for the sake of appearance. Consistency in execution of the procedures has direct bearing on data-evaluation validity. If variability is minimized, the likelihood of detecting a true difference between the groups (e.g., experimental and control) is increased.

Variability of procedures can take another form, particularly in studies where interventions are evaluated. Patient adherence to various conditions can introduce biases that threaten experimental validity. For example, if patients are asked to engage in specific behaviors as part of a treatment program, some will and some will not, depending on the demands made of them. At the end of the treatment trial, there may be no differences among the treatment conditions. This can be due to diffusion of treatment (threat to internal validity). Essentially, no treatment (not adhering to the intervention) diffused or penetrated the varied conditions. Also, the variability in implementation of a given treatment (some carried it out great, others mediocre, and others not at all) is a huge additional threat. The variability is very likely to contribute to a no-difference finding.

Does this ever occur, and is it really important?

Yes on both counts. For example, a study with more than 5,000 women in three African countries received one of three treatments to prevent HIV infection (vaginal gel with antiretroviral drug, a pill with that drug, and a pill combining multiple drugs) (National Institutes of Health [NIH], 2013c). There was a placebo group as well. All patients were counseled how to carry out treatment and received free condoms and ongoing counseling. Bottom line, many individuals in the groups could not adhere to the intervention. There were no group differences in rate of HIV infection that emerged. The three treatment groups were no different from each other or the placebo group. Adherence could be evaluated by checking blood levels of the medication and leftover pills and gel applicators. Poor adherence to the treatments was low (30% in the groups). We do not know about the effectiveness or differential effectiveness of treatment because of adherence to the procedures.

There are many lessons from the example beyond variability of procedures. Treatment trials and then any extension to clinical practice need not only to develop effective treatments but also to be assured that they can be and are carried out.

3.8.4: Unreliability of the Measures

Reliability refers to the extent to which a measure assesses the characteristic of interest in a consistent fashion.

This is a weighty topic we will take up again in the context of assessment. For this chapter, we are concerned about variability in the measure that might constitute a threat to data-evaluation validity. In principle, this threat is similar to subject heterogeneity and variation in procedures; if you have learned one, you have learned them all in the sense that each is about a source of variation (subjects, procedures, measures) that might be better controlled in a study.

Variability in performance on a measure has many sources. One source is the extent to which individuals actually vary on some characteristic of interest (e.g., conscientiousness, warmth). That is not the facet of concern here. Rather, we are concerned with features of the measure that may foster error, inconsistency in responding, and hence unnecessary variability. Performance on the measure may vary widely from item to item within the measure because items are not equally clear or consistent in what they measure and hence performance may vary widely from occasion to occasion. To the extent that the measure is unreliable, a greater portion of the subject's score is due to unsystematic and random variation.

Other variation not of interest for the moment is worth distinguishing nevertheless. Performance is variable from occasion to occasion as a function of mood, experience, context, and many other unspecifiable influences. Thus, even if performance on a measure is perfectly reliable from internal analyses of the scale, as humans we are likely to respond differently to it from one occasion to the next because performance is multiply determined and on any given day our score might well be a little different. Even so, one wants to limit extra, unneeded, and unsystematic variation from the measure. That can be facilitated by using measures that are well studied and known to be consistent in the characteristic(s) they measure. Consistency of performance on the measure may be reflected in many indices of reliability, and these vary as a function of the goals of the measure and the study. Measures that are not very reliable means they have more error, which of course is reflected in the denominator of the ES formula. In studies with relatively unreliable measures, the obtained ES is likely to be lower than it would be if more reliable measures were used. Selection of poorly designed measures in which reliability and validity are in doubt can threaten data-evaluation validity. As unreliability of the measure increases (error), the likelihood of detecting a statistically significant effect increases. Said more blatantly, you tell your office mate, "My results on that measure did not come out." She says, "No wonder, there is no evidence that the measure reliably assesses the characteristic you care about."

It is useful to be wary in one's own research or the research of others about using measures that get at the construct, that are home-made, that have no background evidence in their behalf, and that are used in ways that depart from the uses that have supporting data on the validity of the scale (e.g., with a sample quite different in age from the usual use, of a different ethnicity). More generally, we need to ask at the beginning of the investigation, what is the evidence that the measure assesses the construct of interest in this study (validity) and that the measure does so reliably?

One reason investigators evaluate the reliability (e.g., internal consistency, test–retest correlation) of the measure before the primary data analysis is to see the extent to which they can be assured error was relatively small. Another strategy is to use multiple measures of a construct, check to see that in fact they are related (correlated), and then to combine them statistically. This can be done by placing all measures on a standard score (e.g., mean of 50, standard deviation of 10) and then adding the scores together. Combining multiple and related measures of a given construct can provide a more stable estimate of the characteristic of interest. From a power standpoint, this is a useful strategy.

3.8.5: Restricted Range of the Measures

When groups are compared in an experiment (experimental vs. control group) or observational study (person with a specific characteristic or diagnosis vs. another type of person), we are looking for group differences. When there really are group differences, we might not be able to detect them if the range of the measures is too restricted. Consider the principle to convey the point and then a more realistic example.

Let us say we want to compare adolescents who do or who do not engage in self-injurious behavior (e.g., self-cutting). We believe that these two groups will differ on some cognitive task related to handling stress. We recruit subjects, screen for self-injury, and measure them on cognitive distortion about some topic (e.g., the likely impact of health foods and exercise). As it turns out, there is no measure or what we want to assess so we invent a scale of one item. The one item is scored on a 3-point scale as displayed in Table 3.5.

Table 3.5: Hypothetical One-Item Scale of the Impact of Healthy Foods and Exercise

Score	Result
1	Healthy food/exercise do not really help people
2	Healthy food etc. help people a little
3	Healthy food etc. help a lot

We administer the measure to our groups and find no difference. (Leave aside for a moment that this would be a methodological low if anyone did this in assessment—a one-item measure with no established validity or reliability.) One interpretation is that the groups really are not different on the underlying construct. Another possibility is a threat to data-evaluation validity, namely, the restricted range. A measure with a total possible score of 3 might not spread out the groups sufficiently to show an effect. Most of the people may actually fall into a score of 2 or 3. The restricted range here relates to the numerator of the ES formula. The means of the groups had no place to go to spread out (in range of possible scores), and it would be very difficult to show a difference. The remedy is that we want measures that can range from some low score to some much higher score so that differences can be detected.

Investigators occasionally "throw in" a home-made scale with one or a few items. Usually, there are three problems with such measures:

1. there are no validity data to know what those items measure, no matter what they "seem" to measure;
2. there are no reliability data to suggest that whatever is being measured is done with any consistency; and
3. the very restricted range (variation) of possible scores may interfere with demonstrating group differences when such differences exist in the underlying construct.

The first problem relates to construct validity (we do not know what was really measured); the second and third problems relate to data-evaluation validity (possible variability from a measure with low reliability and hence much error, and restricted range of scores).

3.8.6: Errors in Data Recording, Analysis, and Reporting

A threat to data-evaluation validity obviously would be any facet that could contribute to inaccuracies of the data and their presentation. Several kinds of problems are included, such as making errors in recording or computing the data, analyzing select portions of the data, and fabricating or "fudging" the data. All errors potentially make a difference; some are accidental or careless and otherwise unintended (e.g., miscoding, or misreckoning of variables, such as sex, ethnicity, or group condition). Other errors are intentional (e.g., selective reporting of data, fudging the data). Fudging the data in particular completely rocks the pillars of science as an enterprise and goes way beyond a mere "threat to validity." We will take up later intentional manipulation, alteration, and misrepresentation of the data in detail, but note here the obvious—conclusions from the data can be inaccurate for a variety of reasons.

Unintentional errors in recording or calculating the data include inaccurately perceiving what the subject has done, arithmetic mistakes, errors in transposing data from one format to another (e.g., questionnaires to data sheets or computer files), and similar sources of distortion that can be systematic or unsystematic as shown in Table 3.6.

Table 3.6: Error Types in Data Recording, Analysis, and Reporting

Error Type	Definition	Characteristics
Systematic Errors	This error type means that scores or characteristics of the subjects were miscoded or recoded in the same direction.	Systematic errors in the data may alter the affirmative conclusions.
Unsystematic Errors	This error type means that errors were random or showed no pattern. Each type can serve as a threat to data-evaluation variability.	Unsystematic or random errors in the data may negate or obscure group differences because the errors add variability to the data.

Evaluation of recording and arithmetic errors across several studies has yielded low rates of error, usually hovering below 1%. The heavy reliance on collection via the Internet, laptops, tablets, and smartphones and completion of measures that are automatically scored and go into a database can aid in reducing computational errors.

To be sure, there are obvious advantages in the use of computers in scoring and checking data and computing or transforming scores based on operations that previously would be completed by calculator or by hand. The main advantage is evident when subjects respond directly on a computer or related device (e.g., keyboard, touch screen, tablet, smartphone) and the data are automatically scored and entered on a spread sheet or database. Intervening steps (e.g., scoring the data, entering the data) are reduced or eliminated, along with the opportunity for errors. That said, research is clear here too. The use of laptops or handheld devices to collect data has their own sources of error. In a review of studies using these devices, data recording errors were less than 1% of the data recorded (Haller, Haller, Courvoisier, & Lovis, 2009). This seems small and perhaps we should not worry. Yet we should worry for a couple of reasons.

First, I have only mentioned one type of data error in the above example, namely in recording or coding data. There are other types of errors. For example, a review of 281 articles in psychology journals revealed that 18% of the statistical tests were incorrectly reported, and 15% of articles included at least one statistical conclusion that on recalculation proved to be incorrect (i.e., went from statistically significant to nonsignificant or vice versa) (Bakker & Wicherts, 2011). The point here is that there are multiple opportunities for error in data recording, analysis, and

reporting. Their cumulative impact is unknown, but we already know from the study just cited that the impact of one of these can be huge.

Second, I distinguished intentional and unintentional errors in managing of the data, but it is useful to ignore that distinction for a moment. Errors in data recording and reporting more often than not tend to be in the direction of the investigator's hypotheses. This clearly includes a systematic and directional bias in interpreting results from a study. Thus, the motivation of the investigator (intended or unintended) is not really the point. Errors that appear to be careless or random more often than not are in the direction (support) of the investigator's hypothesis. This not only could be fudging but also could be selective inattention to data problems if the data seem to support what was expected.

Other data analyses threats stem from biased selection on the part of the investigator of those data that should be analyzed or reported. In most studies in clinical psychology, multiple measures are used. These may be different ways of measuring the same construct (e.g., self-report, interview ratings of depression) as well as measures of different constructs (e.g., stress, symptoms of trauma, love of methodology). When the study is completed, the investigator begins the analyses and may find that findings did not "come out" on key measures. That is, the differences were not significant. The investigator may selectively report the results. This can take different forms. First the measures that did not yield significant results may now just be dropped from the study. That is, they will not be reported. Second, what was the primary or main measure of the study may be shifted in light of a look at the statistical analyses. The researcher replaces the original outcome measure with the one or ones that came out to be significant. It is difficult to tell how pervasive such practices are, but there are some data. In one study, planned projects were compared with published reports (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004). The surprising finding; 62% of the 122 studies examined either changed, introduced, or omitted measures. A small subsample of investigators was surveyed, and 86% of the responders (42/49) denied that there were unreported outcomes in their studies despite evidence to the contrary. Other reviewing research has found similar findings with investigators changing, introducing, or omitting measures and selectively reporting significant results (Dwan et al., 2008).

Needless to say, a reader of a given study may not have any idea what the original planned set of measures included, how many were added or dropped, and as part of this how many statistical tests were done to find those that were significant. Selective reporting is a threat to data-evaluation validity because the findings might be seen as very different if all of the data were presented and if the final analyses remained true to the original predictions by keeping the main variables as the main variables.

The problem of omitting data usually rests with the investigator but not always. On more than one occasion, I have been asked by editors to omit measures and data analyses from a manuscript before it is published in the journal. The nonsignificant findings for a set of measures were not interesting or informative and would take up journal space. Whatever the source, the implication is clear, selective use and reporting of data distort the conclusions that would otherwise be drawn.

The selective reporting of data and data analyses raises a broader issue. Many experiments are completed and yield findings that are not statistically significant. We know not only from common sense but also from evaluations of research that studies with significant effects are more likely to be published and when comparisons are made between published and nonpublished studies, the effects that were obtained are larger for the published studies (Hopewell, McDonald, Clarke, & Egger, 2007). In short, there is an unpublished literature and that cannot be neglected. Selective reporting (publication) of findings can have enormous consequences. In the extreme, this is obvious—we have 200 unpublished studies that did not find very much and 2 published studies that had strong effects.

The results of experiments with findings that are not statistically significant are much less likely to be reported than those that attain statistical significance (e.g., Chan & Altman, 2005; Chan et al., 2004). The studies that are not reported are so to speak allocated to a file drawer, if they are written up at all.

> The *file drawer problem* (Rosenthal, 1979), as this is sometimes called, refers to the possibility that the published studies represent a biased sample of all studies that have been completed for a given hypothesis.

Those that are published may be the ones that obtained statistical significance, i.e., 5% at the $p < .05$ level. There may be many more studies, the other 95%, that did not attain significance (Francis, 2012; Pautasso, 2010). The file-drawer problem is a concern across diverse areas of natural, biological, and social sciences as well as research in business and industry. Among the issues is the failure to replicate studies because those that are published do not represent the findings usually obtained. A related issue is overestimation of ESs in meta-analyses because the small or minute sizes are never published. These issues are not a methodological nuance—significant medical findings, for example, may not be replicable because of the publication bias suggesting there is a real effect but because of selecting studies that show that effect (Bakker, van Dijk, Wicherts, 2012).

Methods can be used to estimate how many studies with no-difference findings would be needed to place reasonable doubt on a finding that has attained significance (see Rosenthal, 1984; Scargle, 2000). Thus, the bias can be addressed. For present purposes, the broader point is

critical, namely, findings must be viewed in a broader context of other findings and other studies that attempt to replicate the research. Publication of findings that did not support a hypothesis and that did not show statistically significant effects is unusually difficult unless the negative result addresses a critical issue, shows an exception that has theoretical implications, or provides an intriguing exception.

3.8.7: Multiple Comparisons and Error Rates

Not all of the threats to data-evaluation validity pertain to variability. Statistical evaluation of the results can be hindered by other problems that directly influence whether the investigator concludes that groups differed. In an investigation, many different measures are likely to be used to evaluate the impact of the experimental manipulation. For example, in a treatment study, the clients, clinician, and perhaps relatives of the client are likely to complete a few measures (e.g., depression, symptoms in diverse areas, impairment, and quality of life). At the end of the investigation, experimental and control conditions will be compared statistically on each of the measures.

There are separate but interrelated problems that reflect a threat to data-evaluation validity. The main problem to note at this point pertains to the number of statistical tests that will be completed. The more tests that are performed, the more likely that the difference will be found, even if there are no true differences between conditions. Thus, the investigator may conclude mistakenly that there is a difference between groups and a true effect of the intervention (*Type I error*). The possibility of this occurring is evident in any experiment. The risk of such an error (Type I error) is specified by alpha or the probability level that is used as a criterion for statistical significance. Yet this risk and its probability level apply to an individual test. When there are multiple comparisons, alpha is greater than .05 depending on the number of tests. The risk across several statistical tests, sometimes referred to as *experiment-wise error rate*, is much greater. The number of tests within a study can lead to misleading conclusions about group differences.

The misleading conclusion is a threat to data-evaluation validity. The threat can be exacerbated when investigators conduct scores of tests with varied permutations of the data (e.g., omitting items or subscales of a measure; combining some groups and omitting analyses of some complete measures altogether). If all of the analyses are not reported, the reader has no idea of the extent to which the statistically significant results that are reported could be due to chance (Simmons, Nelson, & Simonsohn, 2011). At this point, it is useful to note that multiple statistical tests serve as a threat to data evaluation.

3.8.8: Misreading or Misinterpreting the Data Analyses

I am hesitant to include this final threat because it will not be credible or look silly. Yet, after all that has been done to design a study, get approval from an Institutional Review Board that reviews the proposal before a study is conducted, recruit and run subjects, the data analyses were done, and what more could happen? This is the last threat to any kind of validity, so let us go so far as to say a wonderful study was done that addressed all threats that have been invented and a couple that have not been. There is now a new threat, namely, the authors' misreading or misinterpretation of their own data analyses. Why is this a threat? Well because the conclusions the authors reach are not the ones to which they are entitled from the data analyses. You might think this could never happen.

Consider a study that compares two groups (experimental manipulation and no manipulation control group). Each group receives a pretest and a posttest with the manipulation sandwiched in the middle. At the end of the study, within-group or correlated t tests are performed. That is, a t test is run to see if the experimental group changed from pre to post. Lo and behold participants in that group did change ($p < .05$). We run that same test for the control group, and they did not change from pre to post ($p < .20$). Now the author concludes that the effect of the manipulation was larger than that of the control procedure and that the conditions are different. It sounds so compelling, but it is a misread of the data. The comparison of primary interest is the comparison of the two groups at post (using repeated measures or analysis of covariance to take into account the pretest). If this is done in an analysis of variance, there would be a Time effect (are both groups any different at pre and post?), a Group effect (are the groups different when summing across pre and post?), and the Time × Group interaction (was one group significantly different from the other at one of the time periods [post]?). The interaction is the test that is needed. Authors making conclusions based on one group changing significantly and the other not without a direct comparison is a misread of what those analyses can yield. If the groups are not different from each other at post and with a between-group comparison, one is not entitled to include they are different. The problems have been stated so well by other authors who make the following recommendation, " . . . when making a comparison between two effects, researchers should report the statistical significance of their difference rather than the difference between their significance levels" (Nieuwenhuis, Forstmann, & Wagenmakers, 2011, p. 1105).

How pervasive could this problem be (and is it just in very low-grade journals where I publish my work)?

Apparently not. A review of over 500 articles on behavioral, systems, and cognitive neuroscience from arguably the

top-ranked journals (*Science*, *Nature*, *Nature Neuroscience*, *Neuron*, and *The Journal of Neuroscience*) found that when the above circumstance (comparisons) were relevant, 78 articles used the correct procedure and 79 used the incorrect procedure (Nieuwenhuis et al., 2011). The authors did a smaller scale replication in a slightly different area of research and found the problem to be even worse. All told, their work covered diverse studies (e.g., when researchers compared the effects of a pharmacological agent vs. placebo; patients vs. controls; one vs. another task condition, brain area or time point; genetically modified vs. wild-type animals; younger vs. older participants). This is one type of statistical issue but enough to be on an alert. Investigators can misinterpret the data analyses and the conclusions they reach. This can misguide readers who may not look closely to see if the proper comparisons were made to justify the conclusion.

There is another data misinterpretation that can mislead. In studies of treatment (psychotherapy, medication, counseling), researchers often compute ESs to complement tests of statistical significance, a practice to be used routinely. The misreading of the data is to interpret ES (magnitude of effect) as a clinically significant effect, i.e., one that makes a difference. A larger ES has no necessary relation to the impact of an intervention on patients in any way that is important to them. We shall take this up separately because the confusion is a common form of data misinterpretation.

3.9: Managing Threats to Data-Evaluation Validity

3.9 Explain the importance of threats to data-evaluation validity in the planning stage

Again, the first step in managing the threats we discussed is to explicitly check the list of threats at the outset when planning a study. It is likely that many of these will be relevant. Low statistical power is likely to be a problem if not attended to directly. It is very easy to check the power of a study before running it.

> Statistical power of an experiment is a function of the criterion for statistical significance (alpha), the size of the sample (N), and the differences that exist between groups (ES).

One can be precise in estimating the sample size one needs to detect a particular level of effect. There are many ways to increase power, and we will take up each of them with concrete recommendations later in the text. At this point, the key issue to remember is that low statistical power often is a threat to data-evaluation validity and more often than not studies are underpowered. By definition that means that if there were an effect of the experimental manipulation (i.e., a real effect in the world), it is unlikely it would be detected in the study.

Subject heterogeneity as a threat to data-evaluation validity focuses on including a broad range of individuals with many diverse characteristics in a study when that is not the purpose. In studies of populations, as in epidemiology and public health, and in longitudinal studies in psychology, sampling individuals from birth through adulthood, capturing a representative sample of the population or all subjects within a particular time frame are important. Yet, the vast majority of experiments in psychology, counseling, and education do not seek representative samples. In any given study, heterogeneity of the sample can mean increased variability and difficulty in demonstrating an effect.

Strategies to manage this threat begin with selecting a homogeneous sample. Homogeneity is a matter of degree. One might wish to limit the age range, type of clinical problem, educational level, and other variables within some reasonable boundaries. Ideally, the decision of what variables to consider and how to limit the variation in the sample is based on theory or research on the effects of these and related variables on the measures of interest. If in doubt, one might select a relatively homogeneous set of subjects as a conservative way of addressing this threat.

A second way to manage subject heterogeneity is to choose heterogeneous samples on purpose but to ensure that the impact or effect of selected subject characteristics can be evaluated statistically in the design. For example, if subjects are recruited for a given psychiatric disorder but some also have other disorders, Co-morbidity could be taken into account in the data analyses by evaluating the effects of treatment separately for cases with and without a co-morbid disorder. More than one variable may be analyzed in this way if it makes sense to do so on conceptual grounds. For example, the data can be analyzed by including subjects of different ages and presence of a co-morbid disorder. In the data analysis, the effects of age (above vs. below the median) and depression (with and without a co-morbid disorder) are included as separate variables (in an analysis of variance or regression analysis). When these factors are analyzed as separate effects, they no longer become within-group or error variance and do not serve to increase the denominator in evaluating treatment differences. For example, in analyses of variance the influence of Co-morbidity (some individuals meet criteria for other disorders but others do not) and all of the other potential sources of variability are in the error term when the *F* test is computed to see if treatment versus control groups (or whatever other groups were included) are different. That means the extraneous variation due to Co-morbidity is in the denominator. By making Co-morbidity a variable in the study (condition [treatment vs. no treatment] × Co-morbidity [yes vs. no co-morbid diagnosis]), this

takes the variability out of the denominator and makes for a more sensitive and powerful test.

In principle and practice, it is possible to analyze the data to death to explore an indefinite set of characteristics that might contribute to the results. Psychological studies typically have too fewer subjects to analyze too many factors, and such fishing expeditions have other problems (increase in the likelihood of chance findings). If a heterogeneous sample is selected, it is useful to begin the study with specific hypotheses about the sub-analyses that will be completed to ensure that these sub-analyses can be conducted with adequate power.

Variability in the procedures can be managed by tightening up the study and how all facets are executed. Potential sources of variation in an experiment include the instructions and experimental material or procedures to which subjects are exposed.

> Variability comes from imprecision in the script or protocol that the experimenter should follow in the experiment. The script refers to the specific activities, tasks, and instructions that the experimenter administers.

Depending upon the investigation, this may entail delivering a rationale, providing a brief interview, answering questions, assisting the subject, and performing a task or implementing the experimental manipulation. The experimenter's script must be well specified by the investigator.

> *Failure to specify in detail the rationale, script, and activities of the experimenter has been referred to as the loose protocol effect (Barber, 1976; Mitchell & Jolley, 2012).*

Several problems may result from failing to specify how the experimenter should behave.

1. First, the lack of specificity of the procedures means that the investigator does not know what actually was done with the subjects and hence cannot convey the procedures to other investigators. The study cannot be repeated either by the original investigator or by others because of the lack of important details.
2. Second is the prospect of inconsistency among different experimenters when two or more experimenters are used to run the experiment. The procedures may vary systematically from experimenter to experimenter in terms of what is said to the subject, the general atmosphere that is provided, and other features.

This variation in experimenter behavior is more likely when details of implementing the procedures are not well specified. Inconsistencies among experimenters may readily obscure the effects of an independent variable. When the experimenters perform differently, this introduces extraneous variability that can dilute the ES.

Standardizing the rationales, procedures, and experimenter's script is a matter of degree. And one ought to decide whether any procedural influence might be better controlled. For example, the study may include running subjects through a scanner (e.g., functional magnetic resonance imaging [fMRI]) as an experimental task is presented. The task and fMRI procedures may be standardized. Yet, could it make a difference what is said to the subject as they spend 2 minutes walking to the scanner or preparing the subject for the actual procedures? The experimenter may vary in topics and emotional tone of those topics or vary in how they are handling reassurance if the subject is concerned. The first question is whether any of this interaction could add error. If so, tighten the protocol to specify what is said, what is not said, and then check to be sure that this is executed correctly.

To ensure that the experimental procedures are conducted in a consistent fashion, the procedures should be explicit and standardized for the experimenters. For laboratory research, and in varying degrees in applied research, many aspects of the procedures can be automated or recorded in advance. Audio or visual recordings of instructions to the subjects, laptop presentation of instructions, tasks, and other material can ensure standardization. When these options are unavailable or seem undesirable by virtue of the goals of the study, the statements to be made by the experimenters may be spelled out verbatim or with strong guidelines. Detailed specification of the rationale or instructions guarantees a certain amount of consistency. Experimenters may vary some of the words used and introduce their own statements, but these do not necessarily compete with the overall consistency of the script.

> Another recommendation is to train experimenters together. During training, experimenters can practice conducting the experiment on each other or the investigator as subjects to see how the procedures are to be performed. By having experimenters practice and receive feedback together, relatively homogeneous behavior during the actual experiment is more readily assured.

Homogeneity in performance can be sustained by conducting training sessions periodically with all experimenters as a group while the experiment is actually being run. One procedure to examine and sustain consistency of performance among experimenters is to include "subjects" in the study who are working for the investigator. These subjects, referred to *confederates*, enter the study as if they were completing the experiment. However, their task is to discuss with the investigator what was done, how it was done, and so on after they participate in the experiment. In my own work, occasionally I have utilized as confederates persons who know the procedures well because of their prior work as experimenters. Perhaps the most useful facet of the procedure is to tell experimenters at the beginning of the project that individuals will be coming through the experiment as subjects. These confederates are unannounced, of

course, and interspersed with other subjects. Probably, the most interesting aspect of this procedure is that it may increase vigilance of the experimenters, as they ponder who is working as a confederate and remain especially careful in adhering to the experimental script.

Finally, experimenters ought to be encouraged to report sessions in which they have deviated from the script. Experimenters should not be expected to perform consistently beyond a certain point and to be entirely free from error. For example, subjects may be run in a condition other than the one to which they were assigned, receive a portion of some other condition, or through some unusual event receive a diffuse or interrupted version of their condition. Ideally, the investigator establishes a climate where high standards of performance are expected yet errors are readily acknowledged and reported to serve the goals of the research, namely, to provide a meticulous test of the hypotheses. Encouraging experimenters to report instances where they inadvertently deviated from the script or were forced to deviate by virtue of the subject's behavior will help the investigator monitor the sorts of inconsistencies that transpire. Gross deviations from the procedures may require excluding subjects from data analysis.

In some studies, it may be difficult to standardize too rigidly what experimenters do. For example, in interventions studies (e.g., treatment, prevention, education), multiple sessions may be provided over time (e.g., weeks, months). Opportunities for loose protocols that increase variability (data evaluation threat) and diffusion of treatment (internal validity threat) are huge. Manuals are often written to dictate what the intervention is, how it is to be conducted, on a session by session basis. Then the fidelity of implementation (treatment integrity) is evaluated, a topic that we will return. Even so flexibility may be essential in response to individual clients and their special situations during the course of treatment.

Managing data-evaluation threats related to measurement involved two issues. The first was unreliability of the measures. Not too much to say here except to have a strong rationale for why a particular measure is used and then data from prior studies or within the study one is conducting to suggest that in fact this was a reliable measure with minimal error. The second one was using measures with established validity. If a measure is not available, some facet of measurement validation ought to be reported within the study in which the measure is first used. Measures of one or a few items are occasionally introduced in research in which the investigator makes up the items to assess a construct of interest. The immediate concern is that rarely is the validity of these established, so what they really measure can be challenged, no matter how intuitive the items seems. For example, how much do you love your uncle might measure love of an uncle but might just as well measure social desirable responding because of the reactivity of the situation. Related to measures with just a few items was the restricted range of scores that are possible. Restricted ranges can limit the ability to show an effect when there might be one. Measures of just a few items can make this problem more likely.

Errors in data recording, analysis, and reporting are weighty threats. The recommendations for managing various biases that may enter into the data vary greatly depending upon the precise source of error. Mis-recording and miscalculating the data are relatively easily controlled, although they may be difficult to eliminate entirely in very large databases. Obviously, individuals who record the data should be kept uninformed of the experimental conditions so that the possibility of directional (biased) errors in favor of the hypotheses is removed. Scoring and entry of the data can include a variety of steps that may vary as a function of the nature of the data, such as whether the dependent measures (e.g., questionnaires) are scored by hand or by computer, whether data are entered directly from scored forms or are first entered on to data sheets, and others.

> Whenever possible, it is preferable to have subjects enter their responses directly on a computer (tablet, keyboard, touch screen, smartphone). The goal is to streamline data collection so that entry by the participant can go directly into a database without research assistants or others involved in data recording or entry.

Errors still can occur. Participants still misread items and indicate their responses incorrectly; software codes to score a given variable may be not quite correct.

It is important to build into the study procedures to check the data closely before analyses.

- If errors are possible (e.g., tapes are scored and the data entered) and human observers or recorders are involved, double score the data, check discrepancies between observers, and check data entry to make sure that every number has been entered correctly.
- Also check the numbers for each dependent measure to look at the obvious.
- Do the data show the correct number of subjects in each condition, on each assessment occasion, for each measure?
- Does the range of scores for each of the measure reflect legitimate scores?
- The measure may have a maximum score of 100, but the range could show that one subject has a score of 200.
- Are there individuals (outliers) whose scores are 2 or 3 standard deviations above or below the mean?
- Is this accurate or a data error?

Compulsive checking may be time-consuming, but it involves a relatively small cost considering the amount of

time that goes into the planning and implementation of the experiment.

If all of the data cannot be checked, certainly a generous proportion from all conditions should be randomly sampled to provide an idea of whether errors occurred and what their influence on the results might be. Checking is important for the obvious reason of detecting and correcting errors. Perhaps as well the checking conveys to all those involved in the research process the importance in accuracy and integrity of the data.

Selective reporting of data and fudging (two practices related to data-evaluation validity) require much further discussion in terms of both the nature of the problems and the range of remedies to manage them. The matter of altering the primary measure or dropping measures from the study in light the results of the statistical analyses has a partial remedy (Chan et al., 2004). When clinical trials are comparing alternative interventions or an intervention against a control group, funding agencies (e.g., National Institutes of Health), organizations (e.g., World Health Organization), and a consortium of journal editors (the International Committee of Medical Journal Editors) require individuals to register their clinical trials in advance of the study. Investigators complete information to convey exactly what the measures are, what the primary measures will be, and how the measures will be examined (see DeAngelis et al., 2005; Laine et al., 2007). And the material is in the public domain. This is an excellent strategy. ClinicalTrials.gov in the United States is the largest clinical trials database, and over 190,000 studies have been registered, encompassing all 50 states in the United States and 190 countries (http://clinicaltrials.gov/). In principle this is excellent as a strategy and one hopes the practice will grow. The issue to keep in mind is that a vast majority of research is not funded at all and does not consist of clinical trials. Yet protecting against selective reporting is important to all kinds of research. The current registry system misses most research studies but might set a standard that is adopted more universally.

The problems of selective reporting go well beyond a threat to data-evaluation validity. They raise issues about training of researchers and inculcating the goals of science and the responsibilities of investigators. (We will take up these reporting issues again in a later discussion of scientific integrity and responsibilities of investigators.) Presumably, instructing investigators about the need to plan analyses in advance, conveying their responsibilities in the reporting of data and their analyses, and noting the consequences of selectively reporting data may help. Yet as I mentioned, publication biases and occasional editorial practices foster selective reporting. Probably one of the best checks is to replicate work that has been reported. This not only addresses the veridical nature of the findings but serves many other functions in the accumulation of scientific knowledge.

3.9.1: General Comments

As obvious from the discussion, several features of the data evaluation can interfere with drawing valid conclusions. The threats related to data evaluation often serve as the tacit downfall of an experiment. Well-controlled experiments that test well-conceived ideas often have weak power, a topic we shall take up further. Perhaps even more pervasive is the hidden variability that can emerge in all facets of experimentation and can obscure differences between conditions.

The notion of experimental control, when first introduced into the discussion of research, is usually raised in the context of control groups and threats to internal validity. However, a deeper understanding of the notion of control stems in part from its relation to data-evaluation validity. The control and evaluation of variability in research, to the extent possible, are critical. The initial question of interest in designing a study is likely to be: Are the groups or conditions different on the dependent measures? The next question is "if there is a difference, will this study be able to detect it?" This latter question raises concerns over data-evaluation validity.

Whether a difference can be detected is influenced by several features of the design (e.g., sample size subject selection) and procedures (e.g., implementation of the intervention, training of the experimenters). Many facets of research including recruitment of subjects, preparation and delivery of experimental instructions, and methods of scoring and checking data all become potential sources of uncontrolled variation and can introduce ambiguity into the results. Error and sloppiness each has its own consequences (as my dissertation committee was overly fond of stating), but they unite in a final common pathway of increased within-group variability. This variability dilutes the obtained ES and diminishes the likelihood of statistical significance when there is a real effect to detect.

We have discussed variability and variation as if it were the enemy. There is some sense in which this might be true, but great care is needed in making this point. The goal of our research is not to eliminate variability but rather to understand it, so it is not quite a coincidence that a basic statistical test that is taught is called "analysis of variance." Conceptually, analysis means we wish to elaborate the full range of factors that influence affect, cognitions, behavior, neurological processes, and domains of interest. These factors include our experimental manipulations interventions (e.g., a new prevention program), those interventions of "nature" not under our experimental control (e.g., childhood experiences, past and present stress, and in general any historical and maturational influence), and individual differences (e.g., temperament, genetic predisposition, and personality style). When any one or more of these serve as the focus of our study, we need to control other sources of variation because

the source of variation that is of interest in our study may be obscured by allowing free fluctuation of all other sources of variation. Research design, various methodological practices, and statistical evaluation are tools to help separate and evaluate these different sources of variation.

One cannot underscore enough the importance of care in conducting a study (e.g., to ensure subjects received the conditions to which they were assigned and correctly and that experimenters or materials presented to the subject render the conditions faithfully). Data recording, analysis, and reporting are part of this. Reaching valid conclusions can be undermined in many ways. The quality of a study entails the entire process from a good idea through the complete write-up and reporting.

3.10: Experimental Precision

3.10 Identify ways to address the problems faced during experiments to obtain the best outcome

We have covered internal, external, construct, and data-evaluation validity. At the design stage, all of the threats to validity ought to be considered. The summary tables in the chapter for each type of validity serve as a checklist of methodological basics to address at the design stage. Not all of the problems that can interfere with valid inferences can be predicted or controlled in advance (e.g., loss of subjects over time). However, most can be addressed in planning the experiment and its execution. Also, even those that cannot be resolved in advance are worth considering at the design stage.

3.10.1: Trade-Offs and Priorities

Addressing each type of validity and all of the constituent threats each encompasses is not possible in a given study. The reason is that addressing one type of validity often compromises another type of validity. That is why decisions in designing a study may involve trade-offs where priority of the investigator determines where the emphasis and methodological controls ought to be.

Perhaps the easiest example to convey potential trade-offs can draw on subject heterogeneity, which was discussed as potential threat to data-evaluation validity. The more heterogeneous the sample, the greater the variability—that is by definition. So let us say we want to study individuals with bipolar disorder and to see how they are different from individuals without the disorder. Just focus on the patient group for a moment. Do we accept all individuals with that disorder into our study—any age, individuals who also have other psychiatric diagnoses, individuals undergoing any treatment, and so on? There is no single or correct answer. We merely need to remain alert to the fact that the less restrictive we are, the more variability in the sample we are allowing. The more variability we are allowing, the less likely we may be to demonstrate an effect. Recall the ES formula and the comments about large denominators. The trade-off is clear. For data-evaluation validity concerns (and statistical analyses), we lean toward selecting a homogenous sample and therefore specifying criteria (inclusion and exclusion) as to whom we allow in the study.

On the other hand, we might like our research to have broad external validity and apply to all or at least most patients with bipolar disorder. Yet, our screening and inclusion criteria made our sample very homogenous and fairly unrepresentative of all individuals with bipolar disorder who often have all those other conditions we used to exclude subjects. So what to do: restrict the sample and possibly sacrifice generality of the findings to the larger population or leave the selection criteria wide open to get a sample very likely to reflect all bipolar patients. When in doubt, err on the side of more homogeneous. The reason: the first task is to provide a strong test of your hypotheses and from the perspective of data-evaluation validity, err on the side of less error variability due to measurement, subject differences, sloppiness, and so on. If the hypotheses are supported, then extend to other samples, settings, and conditions.

The feature to like about methodology is that often many options are available to address threats or specific problems. For example, specific threats or potential problems can be addressed in the design (e.g., selection of control groups), assessment (e.g., add measures that might address a problem such as beliefs, expectations of experimenters, assessment of treatment integrity), and data analyses (e.g., controlling influences statistically). Consider an example, where the authors wanted to study a heterogeneous sample (large variability) but wanted to "control" the variability and possible confounding factors that might be associated with that sample. The goal of this study was to evaluate whether having a diagnosable psychiatric disorder and having colds were related (Adam, Meinlschmidt, & Lieb, 2013). This is a reasonable query in light of enormous evidence showing strong connections between mental and physical illness, how they often go together, and how they affect each other. For example, harsh early environments for children (e.g., exposure to enduring stress, violence) can alter their immune system in permanent ways and lead to poor health outcomes (e.g., greater rates of serious physical disease and earlier than expected death in adulthood) (Krug et al., 2002; Miller & Chen, 2010).

The main finding could be due to all sorts of other influences (constructs). The investigators were concerned about construct validity, i.e., trying to show it was psychiatric disorder that was related to colds and not some other factor associated with that. They also wanted to reduce the variability (subject heterogeneity) by controlling subject

characteristics statistically. Four sources of heterogeneity that might well relate to mental or physical health were assessed (age, gender, and marital and socioeconomic status) and controlled statistically (using each as a covariate to evaluate and remove the impact statistically). This is equivalent to removing the variables from the error term and controlling their impact. So here is a case in which a heterogeneous sample is fine but the authors were sensitive to the problem that without controls a highly variable sample might produce mixed or unclear effects. The results indicated that the relationship between psychiatric diagnosis and colds remained once these other variables were controlled. Does the study answer all questions (e.g., related to construct validity)? No study can do that, and many variables that were not assessed might explain the relation. Yet that is for future research to resolve.

It is useful to consider all threats to validity before a study is designed. Many will be easily dismissed because of the design (e.g., random assignment, use of comparison or control groups). The task is to consider each potential threat, identify those likely to interfere with drawing valid conclusions, and plan on how those can be addressed. Managing the threats is the goal, and that often can be accomplished in many ways, including who will serve as subjects, how many will serve, and how the data will be analyzed. Each of these is for consideration at the early design stage.

In any case in the study, the investigators wanted a large sample representative of individuals from a large community. Representative sample means that this is a maximally heterogeneous sample (age, education, various psychiatric, and possibly physical disorders) with variability (subject heterogeneity) that is large. A fairly large (for psychology) sample of adults (N = 4,022, 18 to 65 years of age) was selected that represented the community. Assessments were made of common colds in the past 12 months (self-report) and psychiatric diagnoses (anxiety, psychoses, substance abuse or dependence, mood) also were assessed. Specific diagnoses were evaluated, but the overall results can be conveyed here to underscore the main point. The presence of a psychiatric disorder was associated with a 44% higher risk of having experienced a cold within the past 12 months. That is psychiatric disorder, and colds were indeed related.

3.10.2: Holding Constant Versus Controlling Sources of Variation

Threats to internal validity generally can be ruled out or made implausible as rival hypotheses by allocating subjects randomly to conditions and controlling potential sources of bias (e.g., instrumentation, attrition) that might arise during the experiment. Yet, in designing experiments, researchers usually are interested in more than ruling out threats to internal validity; they also are interested in providing the most sensitive test of the independent variable possible. Maximizing the likelihood of detecting the relationship raises issues of data-evaluation validity. The investigator wishes to minimize extraneous influences and sources of variation in how subjects respond in the experiment.

Increased precision is achieved by *holding constant* the potential sources of influence on subjects' behavior other than the independent variable. Conditions are held constant if they are identical or very close to that across subjects and experimental conditions. Of course, one cannot realistically expect to implement an experiment in which all conditions are the same except for the independent variable. To cite an obvious problem, all subjects in the study vary because of their differences in genetic make-up, childhood experiences, physical capabilities, intelligence, age, ethnic background, and familiarity with research. Each factor and many others introduce some variation into the experiment in terms of how subjects respond to the intervention.

The manner in which the independent variable is implemented may introduce extraneous variation into the experiment. Ideally, the conditions of administration among subjects within a given condition would not vary at all. Some features of the experimental manipulation might be held constant such as administering instructions or showing materials to the subjects by computers that will not vary. If an experimenter interacts with the subjects, this interaction may vary slightly across different subjects; if several experimenters are used in the study, even greater variation may be introduced. Other extraneous factors of the experiment such as the time of the day, weather, and how the independent variable is implemented all may contribute to sources of variation. These factors can be *controlled* by letting them vary unsystematically across groups.

Control is achieved by dispersing these factors equally across groups by assigning subjects randomly to groups and by running subjects in each condition over the course of the experiment (instead of running all subjects in the experimental condition in the first half of the study and then all subjects in the control condition in the second half of the study). These and other practices eliminate the bias such influences might exert. However, these factors can be *held constant,* which may even be better from the standpoint of demonstrating the relationship between the independent and dependent variables. By reducing or removing sources of variation, a more sensitive (powerful) test of the independent variable is provided.

Critical Thinking Question

In a study, what is the difference between controlling a variable versus holding that variable constant?

Summary and Conclusions: Construct and Data-Evaluation Validity

Construct validity pertains to interpreting the basis for the causal relation between the independent variable (e.g., experimental manipulation, intervention) and the dependent variable (e.g., performance on the measures, outcomes). The investigator may conclude that the experimental manipulation was responsible for group differences, but the study may not permit this conclusion because other factors embedded in the manipulation alone or in combination with the manipulation might account for the findings. Factors that may interfere with or obscure valid inferences about the reason for the effect are threats to construct validity. Major threats include attention and contact with the clients, single operations and narrow stimulus sampling, experimenter expectancies, and cues of the experimental situation.

Data-evaluation validity refers to those aspects of the study that affect the quantitative evaluation and can lead to misleading or false conclusions about the intervention. Several concepts basic to statistical evaluation were mentioned because of their role in data-evaluation validity and statistical significance testing in particular. These concepts included the probability of accepting and rejecting the null hypothesis, the probability of making such decisions when they are false, and ES. Major factors that commonly serve as threats to data-evaluation validity operate by influencing one or more of these concepts and include low statistical power, subject heterogeneity, variability in the procedures of an investigation, unreliability of the measures, restricted range of the measure, and multiple statistical comparisons and their error rates.

All four types of validity including internal, external, construct, and data-evaluation validity need to be considered at the design stage of an investigation. It is not possible in any one experiment to address all threats well or equally well, nor is this necessarily a goal toward which one should strive. Rather, the goal is to address the primary questions of interest in as thorough a fashion as possible so that clear answers can be provided for those specific questions. The threats identify in advance of a study the problems to which one might be alerted and that ought to be addressed as relevant or potentially relevant to the specific study. At the end of that investigation, new questions may emerge or questions about other types of validity may increase in priority.

> The need for further information is not necessarily a flaw, but rather the continued line of inquiry to which an important study invariably leads.

The obstacles in designing experiments emerge not only from the manifold types of validity and their threats, but also from the interrelations of the different types of validity. Factors that address one type of validity might detract from or increase vulnerability to another type of validity. For example, factors that address data-evaluation validity might involve controlling potential sources of variation in relation to the experimental setting, delivery of procedures, and homogeneity of the subjects. In the process of maximizing experiment control and making the most sensitive test of the independent variable, the range of conditions included in the experiment may become increasingly restricted. Restricting the conditions such as the type of subjects or measures and standardization of delivering the intervention or independent variable may commensurately limit the range of conditions to which the final results can be generalized.

In this chapter, we have discussed different types of validity and their threats. The primary purpose has been to describe these threats and how they operate. In remaining chapters, I raise several of these areas again and more concretely and discuss strategies to address threats and to strengthen the inferences drawn from research.

Chapter 3 Quiz: Construct and Data-Evaluation Validity

Chapter 4

Ideas that Begin the Research Process

Learning Objectives

- **4.1** Assess how a research idea or a question forms the basis of a study
- **4.2** Report the different channels that one uses to develop ideas and questions for study
- **4.3** Examine how understanding of the relationship between variables form the basis of a study
- **4.4** Compare moderators, mediators, and mechanisms
- **4.5** Identify characteristics of the full process of translational research
- **4.6** Define theory
- **4.7** Report the relevance and benefits of theory in research
- **4.8** Analyze the causes that make a research idea interesting or important
- **4.9** Report the importance of the right idea for a research project
- **4.10** Review the steps and decision points to follow when progressing from research idea to project
- **4.11** Summarize the steps that lead to a successful research project design

We have now covered a variety of concepts, including threats to validity and various sources of bias that guide thinking when designing, executing, and evaluating research. All that will be critical to keep in mind as we move forward to elaborate research design issues. Yet we begin here with the first step of a research, namely, what will be studied? How and where does one get an idea for an actual study?

Selection of the research focus refers to the idea that serves as the impetus or focus for investigation. The general idea expresses the relation to be studied (e.g., between stress and perception of other people). This idea gets translated to specific hypotheses or predictions of what will happen when certain conditions are varied (e.g., positive cognitions are planted in the subjects, and they will rate their quality of life more highly even though the cognitions are unrelated in actual content). And then the research moves to another level of greater specificity where precisely the hypotheses are tested in very concrete terms. This chapter discusses the initiation of research, sources of ideas, key concepts that often guide research, and the flow from idea to a specific study.

Deciding what to study can seem like or actually be a daunting task and for the obvious reason. All this research has been going on for decades, and now I show up and need to come up with a study that has not been done, is worth doing, and is feasible in my life time or at least in time for the deadline (e.g., graduation, degree). How do I begin to develop the idea for the study?

4.1: Developing the Research Idea

4.1 Assess how a research idea or a question forms the basis of a study

Developing the research idea can be addressed in several ways. This discussion presents the task in different and somewhat overlapping ways, how it can be conceived and approached, and broad types of research that help orient one in selecting questions for study.

In many ways and often without knowing, people already have a pile of ideas suitable for research. These are beliefs about people, social interaction, what controls behavior, what is involved in attraction or repulsion at first

sight, and more. The task is developing the idea for an investigation and bringing it into a scientific paradigm (e.g., theory, hypotheses, concrete procedures to provide a test, control conditions to ensure the results can be interpreted, and so on). We begin with sources of ideas to begin the process of designing a study.

The research investigation begins with an idea or question that serves as the basis of a study. The question may arise from many sources and from efforts to think about a phenomenon in novel ways (see Leong, Schmitt, & Lyons, 2012; McGuire, 1997). Table 4.1 is provided to give a convenient summary of several ways in which the idea for a study emerges and the source of ideas for many studies. The ideas or sources of research are not necessarily independent or exhaustive. They are useful places to begin to see what kinds of ideas can be tested and what the impetus may be for an investigation.

Table 4.1: Selected Sources of Ideas for Studies

Source of Idea	Defined	Hypothetical Empirical Questions
Curiosity	Special interest from observation, belief, experience not necessarily theoretically or empirically driven.	Are musicians (or leaders, psychiatric patients, Nobel Laureates) more sensitive (or eccentric, motivated, clumsy) than nonmusicians (etc.)?
Case Study	Seeing what seems to be a relation among features within an individual and examining whether the relation in fact exists and has any generality.	Does therapy A (which seemed to make this patient better) lead to greater change than no-treatment or some competing treatment? Do people who seem to . . . (love, despise, or both) their parents have similar views toward their children?
Studying Special Populations	Research that isolates a special group for close analysis of characteristics.	What are the cognitions of individuals with depression? Does the presence of a particular personality characteristic predict other characteristics of interest (e.g., later success, dysfunction, and drug use)?
Studying Exceptions	A variant of the above in which a small subpopulation that violates the general rule is identified and investigated or where a particular principle or relationship is likely to depart from the usual one.	What are the characteristics of children who are abused (or who come from seemingly horrible environments, or who eat horribly unhealthful foods) and have wonderful life outcomes (or experience no deleterious effects)? Or what are the characteristics of people who come from seemingly ideal nurturing environments and have disastrous outcomes?
Studying Subtypes	Also a variant of the above but one in which an overall group that has been studied is evaluated to predict critical distinctions or subtypes.	Can one distinguish in meaningful ways those individuals who are clinically depressed (or who show agoraphobia, high levels of achievement)?
Questions Stimulated by Prior Research	Addressing a question stimulated or unresolved by a specific prior study or area of research.	Studies that identify something not addressed in a prior study. This could be a methodological limitation or competing constructs that might explain the results different from the interpretation by the original investigators. Can a competing interpretation be provided that better accounts for the original finding and makes new predictions?
Extensions of Prior Work to New Populations, Problems, and Outcomes	Efforts to see if the relation affects other areas of functioning or domains not originally studied.	Studies to see if other areas are influenced or affected. Psychotherapy alters symptoms of adults (e.g., anxiety); does the therapy also affect the marital relations or child–parent contacts of the treated patients? Treatment A helps depression; can it also be used for eating disorders?
Extensions of Concepts or Theory to New Problems	Efforts to see if a construct (e.g., addiction, dependence) can be extended to areas where it has not been applied.	Studies that see if addictive behaviors extend beyond the usual use of that term; Is there reward value in aggressive activity similar to the reward value of food, water, or sex?
Extending External Validity	Efforts to see if the relation applies to new populations, settings, and context.	Does the prior finding or theory apply to a different ethnic group or under varied circumstances? Can the treatment be delivered by . . . (parents, college students, computer)?
Translating and Extending from Human to and from Nonhuman Animal Research	Drawing from findings on basic processes or patterns of functioning.	Can exposure to anxiety-provoking stimuli (flooding in animal research) be used to develop parallel treatment for anxiety among adults? Are there parallels in courtship (or communication, dominance, and interactions with newborns) between a specific mammal species and humans, or does the animal research lead to a prediction in one of these areas?
Measurement Development and Evaluation	Efforts to assess a concept (e.g., self-esteem, anger) and to evaluate aspects of the measure.	Studies of the reliability and validity of the measure; utility of a measure in predicting an outcome.

4.2: Sources of Ideas for Study

4.2 Report the different channels that one uses to develop ideas and questions for study

The ideas or sources of research are not necessarily independent or exhaustive. They are useful places to begin to see what kinds of ideas can be tested and what the impetus may be for an investigation.

4.2.1: Curiosity

Many ideas arise out of *simple curiosity about a phenomenon*. This is not a formal way of generating an idea, but it certainly is one to note explicitly.

> Curiosity is no explanation of why a particular course of research is pursued, but it helps convey that the motive for asking particular questions in the context of experimentation need not always germinate out of complex or highly sophisticated theoretical notions.

This research may seek to describe how people are or how they will perform in a particular situation. The more the study seeks to generate and test novel ideas about why people behave in a particular way the better for research, but just beginning with a demonstration that they do or do not behave in a particular way may be interesting by itself.

In many ways, curiosity is an overarching concept that entails other sources of ideas we cover next. In psychology, we do not want a collection of mere associations (correlates) and hence showing that this long list of variables is related to another list of variables may or may not be of great interest. Yet if it is of interest to you, definitely pursue it.

Curiosity may lead to describing relations among variables that were not recognized and then serve as a basis for generating **theory** (why in the world are these constructs related) and then further tests of theory that elaborate that nature of that relation more deeply.

4.2.2: The Case Study

The case study is a special case where curiosity may be peaked and generate ideas for research.

> The case study refers to the intensive study of the individual.

However, this could be an individual person, group, institution (e.g., political body), or society. That these are "cases" pertains to the intensive focus on one or a few instances. The case study has had a major role historically in psychology in prompting interesting theory and research and hence is a valuable source of ideas to consider (e.g., Rolls, 2010). By case study, I am referring primarily to the anecdotal case study in which the assessment is not likely to be systematic and in which control conditions are not invoked. Hence, valid inferences (in which threats to internal validity are controlled) usually are not possible. (I mention this because later we will discuss single-case experimental designs. These designs are experiments that can focus on individuals or group.)

> In clinical psychology and other mental health professions, the case study focus usually is on the individual client, often in the context of the development or treatment of clinical dysfunction. Information is reported about the case that is based on anecdotal information, i.e., unsystematic measurement that is difficult to replicate or verify.

A clinician or client recounts experiences and places the information together in a cohesive narrative that explains something like how a clinical problem came about, why the individual is like he or she is, why and how treatment worked, and similar issues.

For example, recall the case from the 1880s in which Joseph Breuer (1842–1925), a Viennese physician and collaborator of Sigmund Freud (1856–1939), treated Anna O. (Breuer & Freud, 1957). Anna was 21-years old at the time and had several symptoms, including paralysis and loss of sensitivity of the limbs, lapses in awareness, distortions of sight and speech, headaches, and a persistent nervous cough. These symptoms were considered to be due to anxiety rather than to medical or physical problems. Breuer visited Anna regularly to provide treatment. This included talking with Anna and hypnosis. Anna recalled early events in her past and discussed the circumstances associated with the onset of each symptom. As these recollections were made, the symptoms disappeared.

> This case has had enormous effect and is credited with marking the beginning of the "talking cure" and cathartic method of psychotherapy. The report sounds rather amazing and understandably, even with these brief comments, provokes many questions for research.

Before we leap too far, a little dose of methodology and science is important. We have no really systematic information about the case, what happened, and whether and when the symptoms really changed. Also, the case is odd as the basis for the arguing for the effectiveness of talk therapy. For one, talk therapy was combined with hypnosis (which I mentioned) and rather heavy doses of medication (which I did not mention). A sleep-inducing agent (chloral hydrate) was used on several occasions and when talk did not seem to work (see Dawes, 1994). Thus, the therapy was hardly just talk and indeed whether talk had any impact cannot really be discerned. Also, the outcome of Anna O, including her subsequent hospitalization in light of her clinical dysfunctions, raises clear questions about the effectiveness of the combined talk-hypnosis-medication treatment. Cases such as these, while powerful, engaging, and persuasive do

not permit inferences about what happened and why. In fact, most threats to internal, external, construct, and data-evaluation validity apply and are "wrong" with the case. I just hinted at construct validity (talk therapy or multiple treatments that include visits, talk, hypnosis, and medication). But we do not even get to the threats without evidence that in fact there was a change. Yet, we are talking about cases as a source of ideas and hypotheses, and Anna and cases like that raise fascinating questions to be tested.

In psychology there are many other cases where special circumstances such as injury have led to important insights followed by research. Let me provide a case less familiar than Anna O. and focus on brain and behavior. In this case, a 25-year-old man had a stroke, and assessment revealed that he had damage to a specific areas of the brain (insula and putamen) suspected to be responsible for the emotion of disgust (Calder, Keane, Manes, Antoun, & Young, 2000). The damage could be carefully documented (by fMRI [functional magnetic resonance imaging]). His damage could be located to these areas. The man was systematically tested during which he observed photos of people experiencing different emotions (happiness, fear, anger, sadness, and surprise). He had no difficulty identifying these emotions. However, he could not identify the photos of disgust. Disgusting photos or ideas presented to him (e.g., such as friends who change underwear once a week or feces-shaped chocolate [remember, I am just the messenger here I am not making this up]) were also difficult to identify as disgusting. This is an interesting example because the case was systematically evaluated, and hence the strengths of the inferences are commensurately increased. Also, the investigators compared this case to male and female control subjects without brain injury to provide a baseline on each of the tasks. The demonstration becomes even more interesting by falling somewhere between a case study and a quasi-controlled study. Also, the distinguishing feature is systematic assessment so that alone is a leap from anecdotal case studies such as the example of Anna. In addition, the case was used to explore a hypothesis.

Our discussion is not about case studies per se but the use of cases—contact with individuals who have had special experiences—as a source of ideas for research. You see a possible connection (correlation) or observe a couple of cases and see similar connections. Cognitive heuristics and other limitations of our raw and "normal" observations can obscure relations. That means experience by itself is not usually a good test of a hypothesis, but we are talking about sources of ideas and cases can be quite helpful in thinking creatively about correlates, risk factors, and other facets worth studying.

> In general, close contact with individual cases provides unique information because of observation of many variables, their interactions over time, and views about the bases of personality and behavior.

Indeed, in clinical psychology one rationale for practical clinical experience during training (e.g., practicum experience at a clinic, internship) is that better understanding of clinical cases will improve the research a person does, for those who enter research careers. Close interaction with cases might raise questions, such as do most children with autism spectrum disorder (ASD) show this or that characteristic, among couples who are very happy, and do they show this or that characteristic? Cases can generate many hypotheses about all facets of functioning (e.g., impact of special events in childhood, why one relates to others in particular ways).

4.2.3: Study of Special Populations

The *study of special populations* is encompassed by a few of the entries in Table 4.1 (study of special populations, exceptions, subtypes, extending external validity). A great deal of research focuses on a special group of individuals and compares them with others who do not have the special status. Common among such studies are comparisons of individuals with and without a particular clinical disorder (e.g., depression vs. no disorder or some other disorder, who lived or did not live in foster care) or the search for subtypes among all individuals who might be designated as having psychological dysfunction or disorder. A particular clinical problem (e.g., posttraumatic stress disorder [PTSD]), style of functioning (e.g., risk taking), or population (e.g., first-born children, spouses who are violent with each other) may be of interest, and the investigator asks, what are the key and associated characteristics or how do individuals with the characteristic differ from those without the characteristic? The correlates (e.g., in personality style, family background) and similarities and differences among varied clinical problems encompass a wide range of investigations. The special population might be selected because of a particular experience in their past (e.g., sexual abuse, exposure to violence, being an orphan, last born child) or because of a current experience (e.g., victim of trauma such as a natural disaster, becoming a new parent).

A variation of special populations is worth distinguishing and is noted in the table as the *study of exceptions*. We expect or indeed know from prior research that individuals with some experiences or exposure to some factors to have a particular outcome, but there might be exceptions. For example, among soldiers deployed in combat, most do not develop symptoms of posttraumatic disorder, but certainly some do and it is not merely a matter of the trauma experiences to which they are exposed.

Can we study the exceptions, i.e., the many but not the majority who experience trauma?

Perhaps there is something we can identify about them that would allow early identification or even better prevention. One vulnerability factor that can be identified is

higher emotional reactivity, a physiological reaction to some provoking stimulus tested in laboratory experiments. Individuals who are more reactive physiologically are more vulnerable to PTSD in war (Telch, Rosenfield, Lee, & Pai, 2012). More work is needed, and a great deal has been done, but this is an important beginning to elaborate vulnerability (risk factors) and potentially leading to preventive efforts (to alter reactivity among those who might be especially vulnerable).

More generally, the study of exceptions might entail any group of exceptions. For example, people exposed to difficult or horrible experiences (e.g., sexual and physical abuse, extreme poverty) or adversity (e.g., their parents were criminals, alcoholics, or clinically depressed) often function quite well in everyday life.

What are the factors that have protected them from untoward outcomes?

One can see the implied hope the exceptions provide. Perhaps if we understood how individuals fared well in the face of adversity, we could help the many who do not fare so well. Consider a more concrete example: some small number of individuals who contract HIV do not contract AIDS. This suggests that if we could identify the how and why of this phenomenon, we might be able to use that information to protect all or most people.

4.2.4: Additional Information Regarding Special Populations

The examples suggest some untoward experience that does not invariably lead to an untoward outcome. Think about your own situation. You were deprived of methodology early in life and you are still doing all right! Of course, the opposite line of work in studying exceptions is no less valuable. People exposed to seemingly nurturing conditions (high levels of warmth and involvement of both parents, wonderful sibling relations, opportunities and early competencies early in life, methodology bedtime stories every night) may turn out with very difficult lives. In adulthood, they may turn to lives of crime and drugs. What "went wrong?" "Wrong" is not a useful scientific concept per se (we do not deal with right and wrong or the judgments they entail), but the concept is meaningful by asking what accounts for individuals with a particular type of experience (in this instance seemingly close to ideal childrearing) go down one path (functioning well) versus another (not functioning so well). Can research begin to identify what these exceptions are like? Now you develop a hypothesis of who those exceptions are, how they might be identified, and what makes them different. This could be one study, but it could also be a career.

Subjects who are rare exceptions emerge in another context in methodology. They are often referred to as "outliers" and raise issues for data evaluation and statistical analysis.

> Outliers refer to individuals whose scores or performances on measures depart greatly from the rest of the sample.

Occasionally subjects are deleted from the study, a topic we have much more to say about. Yet in this chapter, the study of exceptions has a different thrust. Identify exceptions and study them. This can be extremely important. Also, when interventions fail, there is increased interest in going well beyond what the group showed as a whole.

Are there exceptions, and can one utilize those for greater insights?

For example, most treatments for cancer that make it to the point where they are tested on humans do not help enough people and are no longer pursued as treatments. However, occasionally there are "exceptional responders" to these drugs (Kaiser, 2013b, p. 263). These are individuals who in fact respond extremely well (e.g., tumors are gone and the effects are maintained) even though the treatment did not help most people in the group. Studying these individuals can lead to great insights about tumors and their treatment.

What is it about these exceptions that made them respond well to a treatment that was ineffective for most people?

Some factor must work in conjunction with that otherwise ineffective treatment to make it very effective. In this example, a genetic variation was found in the tumor that characterized the exception. The treatment was then applied to others with that variation, and treatment was effective (Iyer et al., 2012). Without that factor treatment did not work very well, and with that factor it worked extremely well. This is a huge finding. We almost threw way an effective treatment because most people in the group did not respond. We still need a treatment for those individuals of course. Yet, studying exceptions yielded important insights that affect many people. We can now direct individuals to treatments from which they are likely to profit and perhaps by identifying factors that may be altered to make more individuals responsive to treatment. More generally, the study of exceptions can greatly advance our understanding of underlying processes that relate to the unexpected and also expected outcomes.

> Another variation of studying exceptions or special groups focuses on grouping individuals into various *subtypes or variations of a problem*. This begins with interest in a group (e.g., individuals have a particular condition such a depression) and considering possible subgroups based on clinical experience, a hunch, or theory.

Any one of those might pose that individuals with major depression are not homogeneous but include many subgroups. Distinguishing subgroups might be very important in relation to clinical course or prevention. Here the goal of research is to show that there are subtypes and that unique characteristics of the subtypes (i.e., correlates, risk factors) vary.

For example, many children are victims of bullying. They are the object of verbal and physical acts of aggression and intimidation by others. Yet, among victims one can distinguish those who are so to speak "pure" victims and those who are victims/bullies. That is, among victims a subtype can be identified who also engage in bullying.

> The value of identifying subtypes comes from showing that the distinction is important in some way. We have learned that indeed it is. Victims are at risk for all sorts of problems related to mental health (e.g., anxiety, depression, withdrawn as well as disruptive behavior), physical health (e.g., sleep disturbances, stomach aches, and vomiting as a stress reaction), and poor school functioning (e.g., increased absenteeism, decreases in achievement). For those victims who are also bullies, these characteristics and long-term outcome are much worse!

In addition to the victim characteristics, they are overwhelmingly rejected by their peers, and among the groups (bullies, victims, victims/bullies) they do the worse in school (Jimerson, Swearer, & Espelage, 2009). In other words, subtyping here makes an important difference.

Identifying subtypes is an important focus of research in part because the results can have broad implications. If there are, say, two different subtypes of a problem, this might be quite useful in preventing or treating the problem. The different subtypes may suggest different causal paths and allow one to target the intervention to influences that will make a difference for one, both, or more of the subtypes. We will talk about moderators later in the chapter, and moderators are variables that can help evaluate different paths and subtypes or at least variables that change the relationship between one event and another.

4.2.5: Stimulated by Other Studies

A very large portion of the published research is directed at building upon, expanding, or explaining the results of other studies. A broad category for source of ideas then is research *stimulated by other studies*. This is encompassed by a few other sources of ideas, including resolving a specific issue from prior research, extending the focus (outcomes, dependent variables), and external validity (e.g., populations, settings). There is overlap among these, but the emphasis is one that guides different types of studies.

The research stimulated by prior studies may focus on empirical or conceptual extensions to *new populations, problems, and outcomes*. One type of work may extend a given finding to a new or different population or clinical problem. One sees this in drug studies quite often, i.e., if a drug treats depression effectively, can it also be used for anxiety, or eating disorders? For psychological treatments as well, this thinking has been applied. We have thought that treatments are specific to types of problems (use treatment x for anxiety, treatment y for depression, and so on), but many different problem domains (psychiatric diagnoses) are not so distinct (overlapping symptoms, comorbidity) and many treatments are transdiagnostic (i.e., can be extended to more than one disorder or clinical problem and be effective) (e.g., Farchione et al., 2012; Maliken & Katz, 2013). Extending an intervention to new domains (different clinical problems or to consequences beyond those originally studied) is one variation that can be an interesting line of work.

Extending a given finding to a new set of dependent variables or outcomes can be an interesting source of research ideas too. For example, we have known for decades that cigarette smoking increases the risk of lung cancer. Research extended the evaluation of smoking to many other outcomes by showing that smoking increases the risks of many other diseases (e.g., other types of cancer, heart disease) and has impact on nonsmokers and their disease risk if they are in contact with smokers (secondary smoking). And even most recently, the findings have been extended further. From a meta-analysis of 26 studies following smokers up to 9 years, we have learned that individuals who quit smoking, compared with those who do not, have many mental health benefits, including reduced level of depression, anxiety, and stress and improved quality of live and positive emotions (Taylor et al., 2014). Apart from further commentary on smoking, the research in this review conveys the extension of initial findings (e.g., smoking and lung cancer) on a variety of different outcomes.

- *Extensions can be conceptual* in the sense that a model is extended. For example, the use of illicit drugs is often considered to be addictive. To call something an addiction usually means the substance or activities lead to feelings of pleasure and changes in affect and cognitions.
- The substance or activity that triggers addiction must initially cause feelings of pleasure and changes in emotion or mood.
- There may be a tolerance that develops so larger doses or portions are needed to achieve the benefits.
- Withdrawal symptoms (e.g., physical, emotional) may result from withdrawal.
- There is a dependence on the substance or activity in the sense that it is heavily sought and often on the mind of the person who is addictive. Changes in the brain also result along with the changes in affect, behavior, and cognition. Addictions are associated with impairment in performance, such as meeting expectations in one's work, school, and relationships.
- A key feature is difficulty in stopping whether it is taking the substance or not engaging in the activity.

Now with these key features, one line of research is to extend this to areas of functioning not usually conceived as "addictive." For example, is there the equivalent of addiction to social media (e.g., textbook, Facebook), sex, or food? Can individuals be identified who show signs of addiction

in relation to social substances, so to speak, and what are the brain centers involved for these individuals? This research is an extension to see of core features of addiction as studied in a familiar domain (substance use) and extends to domains not usually considered to be relevant. This is different from lay concepts that are extended (e.g., workaholic), which refers to work efforts that are considered by someone to be obsessive or excessive.[1] Rather, the science feature would be to see if in fact for some people "work" operates like an addiction and if so what new do we learn from extending the conceptual view of addiction. We do not merely want to call everything an addiction just because people do it a lot.

Another type of work stimulated by other studies focuses on the interpretation of the original finding. Some original finding is obtained, and the investigator provides an interpretation (e.g., cognitions changed and that is why this worked). You read the study and challenge the interpretation (construct validity) and now consider study pretty much like the original but one in which whether or not cognitions changed or could explain the findings is tested. Perhaps you have another explanation or interpretation (expectations on the part of the subjects) and measure that to test whether the original view (cognitive changes) really explains the finding.

4.2.6: Translations and Extensions between Human and Nonhuman Animals

One way is *extending or translating findings from nonhuman animal research to a clinical phenomenon*. This is not necessarily a test of generality to see if a finding with college students applies equally to centipedes. Typically such extensions focus on efforts to understanding basic processes or mechanisms of action.

In psychology and basic sciences generally, extensions from animal to human research move in both directions. For example, human studies of lead poisoning and cigarette smoking were elaborated by animal studies looking at processes and mechanisms that could explain how these toxins damaged various organs (e.g., dendrite formation in the brain, lungs, respectively). In clinical psychology, more pertinent is the extrapolation of findings from animal research to human behavior as a basis for a study.

Can some process related to development, social interaction, parent–child interaction, and conflict resolution demonstrated in basic animal research be used to inform and to study human interaction?

Of course, to extend animal research to humans does not mean or imply that there are no unique features of a particular species (i.e., us). Also, sometimes the public is loath to learn of continuities if there is the implication that we are "no different from animals." This latter implication is rarely if ever the research agenda. Continuities and discontinuities are important to demonstrate and understand because they have broad biological, psychological, and social implications. So, for example, we know now that dolphins in the wild seem to call (signal) each other by name, i.e., they have names sort of like us (Janik, 2000), that elephants seem to communicate by producing sounds (through the air) but also by vibrations through the ground from foot stomping (O'Connell-Rodwell, Arnason, & Hart, 2000), and that whales teach each other how to hunt through social imitation much like how we learn many behavioral patterns (Allen, Weinrich, Hoppitt, & Rendell, 2013). These findings are not merely interesting but have fascinating implications in relation to language, brain development, and socialization that may transcend any particular species. Research that draws these connections can be extremely informative because much can be brought to bear in understanding by showing the ways in which species are and are not similar. As an example, a huge area of research is nonhuman animal cognition and among the many goals is to understand decision making, planning, modeling (learning by observation), and choice and to extrapolate that to inform human cognition in the process. The underpinnings of cognition can be evaluated by isolating processes more readily than might be allowed in human laboratory studies.

> Many lines of nonhuman animal research (on classical conditioning, avoidance learnings) have ended up generating research on treatments for humans. For example, researchers in Pavlov's laboratory identified a situation in which animals became very anxious when making a difficult discrimination.
>
> The "breakdown" was referred to as *experimental neurosis* (neurosis once was the word for anxiety and anxiety disorder). Decades of research used laboratory-induced anxiety reactions to develop effective treatments for humans.

Extending findings in ways I have discussed are only samples of the range of possibilities. That is why in Table 4.1 another source of ideas is *extending external validity* or generality of a finding to novel populations or outcomes. These are just two avenues of testing the limits of external validity. One cannot list them all or all that would be of interest for a given finding. Yet, it may be interesting to extend a finding to different cultural or age groups, groups with different gender identity or experience, and so on. One needs a strong rationale for this type of research rather than saying I am doing this study because the finding (intervention, independent variable) has never been tried with that group. That rationale alone is usually regarded as very weak in part because there are an infinite number of conditions, contexts, populations, age groups, and so on alone and in combination that could be studied. Because the extensions are limitless, any research is advised to convey why anyone is of special interest, i.e., are there compelling reasons beyond, "this has not been done before."

4.2.7: Measurement Development and Validation

A considerable amount of research focuses on *development or validation of measures*. Developing assessment devices is central because measurement is a precondition for other research. An investigator may be interested in studying empathy, risk taking, hopelessness, adjustment, psychopathology, love, bereavement, altruism, propensity toward violence, extraversion, and so on. As psychologists we are interested in a vast range of constructs and how they operate with many different populations.

Research is begun to develop a new measure and to establish various types of reliability and validity of the measure. In the process of this research, the relations of the measure and underlying construct to other domains of functioning are elaborated. Measurement development is not a matter of listing a bunch of items and having subjects complete them. There are multiple steps. It is important to note here that development and evaluation of measures are the major source of ideas for research projects. In clinical psychology, the presence of a number of journals developed to assessment attests to the importance of assessment issues as a line of research.[2]

4.3: Investigating How Two (or more) Variables Relate to Each Other

4.3 Examine how understanding of the relationship between variables form the basis of a study

There is another way to help with source of ideas for a study. Consider for a moment that the overall goal of research is to understand a phenomenon of interest; that is, we want to know its characteristics, the factors with which it is associated, how it operates, and how it can be controlled. Sometimes the goal of research is stated to identify causal relations, and that is a useful point of departure. Once causal relations are known, we know a great deal. However, there is more to know about relations among variables than their causal connection and also a great deal of important information to know even if we do not yet know about cause. There are many ways variables can relate to each other, and identifying these and the key concepts they reflect also serve as the bases for doing a study.

Consider for a moment that we want to understand the relation of two variables (e.g., substance use and participation in organized athletics). There are all sorts of connections between variables that can be studied. It is useful to organize the discussion by describing and explaining the phenomena we wish to study. For the moment, consider description as the "what" and explanation as the "how." One focuses on what the relation is to some other characteristic; the other is how or why there is a relationship and through what processes they are connected.

> A source of ideas for research is considering what facet of description or explanation might be studied.

Several key concepts serve as a guide to descriptive and explanatory research. Table 4.2 presents key questions and concepts that pertain to the relations among variables of interest and that often serve as the impetus for an investigation.

4.3.1: Association or Correlation between Variables

Research in clinical, counseling, and educational and other areas of psychology often focuses on identifying whether two variables are *correlated*. (We take up the simple case of correlating two variables, but multiple variables can be examined for their correlations.) Subjects are tested on several measures at a particular point in time to relate such variables as symptoms (e.g., depression, anxiety), cognitive processes, personality, stress, family functioning, or physical health, and correlations predicted from theory or another source are examined. Identifying characteristics of affect, cognition, behavior, and the contextual environment (e.g., characteristics of others) that are or are not correlated with a particular problem can be important for elaborating the nature of a problem and for testing or developing theories about the onset or course of a problem.

For example, there is a relation between temperature (in the weather) and violence. In one correlational study, violent crimes (e.g., assaults, sexual assaults, homicide) were assessed along with daily temperature from records obtained over a 7-year period in one city (Dallas, Texas) (Gamble & Hess, 2012). Incidents of violence were positively correlated with temperature. The higher the temperature, the more violent crime but the positive relation only held up to 80 degrees (F). The correlation became negative after 90 degrees. The authors surmised that at higher temperatures people just try to stay in their homes more and hence decrease all activity.

A correlation between temperature and crime is very interesting indeed and prompts questions that, when studied, could move toward a deeper level of understanding. The obvious general questions of course are:

- Why is there a relation and precisely why does that relation turn from positive (as temperature goes up) but then shifts to negative at a higher temperature?
- Is it really temperature or some other variable that is associated with temperature change? (It is difficult to come up with something.)

- Also, the range of temperatures in any given city is restricted; would the relation hold in different latitudes (e.g., cold climates where the temperature may vary but be colder overall)?

You no doubt can generate your own set of questions, but one can see how a correlation usually is a beginning. Now we need to identify possible explanations or a little theory of why the relation holds and test what may be involved with a little more precision. The correlation is a description of a relation, and we would like to move further along toward explanation by ruling out some explanations and making others more plausible.

As another example of correlation of handedness and mental disorder, does being left- or right-handed have any bearing on rates of mental disorder?

Brain structure and function vary by the different sides of the brain (laterality of the brain) and hand dominance relates to that. Also, being left rather than right-handed is slightly associated with higher rates of autism and epilepsy, so associations like this are already known. In a study with individuals being seen for mood disorders or schizophrenia, investigators found no association of left-handedness with mood disorders (Webb et al., 2013). About 11% of the sample showed left-handedness in keeping with the general population. Yet, individuals with a diagnosis of schizophrenia showed a 40% rate. This is quite a remarkable difference—clearly in this study, handedness was associated with a diagnosis of schizophrenia. Now the work begins—we need some testable theory about this—what of brain structure, function, and activity might be a link. How do these characteristics come together? Maybe it is not handedness at all. Left-handed individuals are slightly more likely to have been born prematurely. Perhaps we ought to assess and control (e.g., match samples) on premature birth and rule out that influence. Correlation is an intriguing starting point.

4.3.2: Concepts That Serve as the Impetus for Research

While it is very useful to raise the notion of correlation as a type of focus for a study, one does not merely select variables randomly. A study begins with a reason to pursue a particular correlation. The impetus ought to begin with a view, theory, or interesting question. Table 4.2 lists some concepts that serve as the impetus for research. That said, as an investigator it is our task to show how the correlation might well be of interest or significance and that depends not only on available information (other research of any kind) but also on our ability to put this together in a persuasive and coherent way. We will talk more on that later in the chapter when we discuss what makes a study important.

Table 4.2: Concepts That Serve as the Impetus for Research

Concept	Description
Correlate	The two (or more) variables are associated at a given point in time in which there is no direct evidence that one variable precedes the other.
Risk factor	A characteristic or variable that is an antecedent to and increases the likelihood of an outcome of interest. A "correlate" in which the time sequence is established.
Protective factor	A characteristic or variable that prevents or reduces the likelihood of a deleterious outcome. Time line not always established; these are often correlations that are negative with some outcome.
Cause	One variable influences, either directly or through other variables, the appearance of the outcome. Changing one variable is shown to lead to a change in another variable (outcome).

4.3.3: Risk Factor

Correlation as previously pertains to two (or more) variables that are related at a given point in time. That is, measures are taken at the same time (e.g., as if we assessed a person's height and weight on the same day and that for many individuals). We would have the correlation of weight and height.

A risk factor is a predictor of some later outcome.

A risk factor reflects a deeper level of understanding than a simple correlation because now we have the time line established between the two variables (some event or characteristic) and a later outcome.

In other words, risk factor is a correlation where we know that one variable comes before the other. That is, an experience, variable, or event (e.g., abuse, exposure to religion) is correlated with a characteristic that emerges at a later point in time (e.g., marital happiness).

Risk factor, as a concept, emerged from public health and in the context of studying disease (morbidity) and death (mortality). The term and common foci in that context refer to "risky" practices (e.g., eating high fat diets, cigarette smoking, not taking methodology courses) and deleterious outcomes (e.g., heart disease, death, despair, respectively). However, the term refers more broadly to events, experiences, or practices that increase a particular outcome of interest. The experiences (e.g., meditating, exercising) and the outcomes (e.g., coping well with stress, donating to charity) can be quite positive without being "risky" in any negative sense. Consequently, the term is used to reflect characteristics that are correlated with and antecedent to a later outcome, no matter what that outcome is. Psychologists often avoid "risk" by

talking about a "predictor" of some other variable. A predictor can avoid the awkwardness of a risk for some great outcome.

4.3.4: Understanding the Difference between a Correlate and a Risk Factor

The difference between a correlate and a risk factor is critically important. It is often the case that researchers will identify a sample with a particular problem or clinical focus (e.g., engaging in self-injurious behavior) and at the same time administer measures of other characteristics (e.g., exposure to trauma in the present or past, level of anxiety or depression, personality, social support or friendships) and then correlate these latter characteristics with self-injury. These are concurrent studies where only correlates can be identified. Yet, risk factor may be used to imply that some characteristics antedated self-injury and only a longitudinal design can unequivocally establish the time line.

Risk factors are not to be confused with a cause, although they so often are. For example, risk factors for heart disease include elevated cholesterol, cigarette smoking, lack of exercise, being short, bald, and male, to mention a few. None of these necessarily causes a heart attack, although all combine to increase risk. With a risk factor, we know that some early experience or exposure, for whatever reason, increases the likelihood that the later outcome will occur. In contrast, demonstrating cause means, of course, that we have established the relation is not merely in a temporal ordering of events but rather some direct influence. I shall return to cause but worth mentioning now is that one can move from risk factor to cause as a research focus. Some risk factors can be altered (e.g., harsh parenting) by changing the environment or person in some way.

An excellent idea for research is asking "If this risk factor is changed, would there be a change in the outcome?" That moves to causality.

A recent example of a risk factor in the context of autism has been intriguing. The placenta is the organ that nourishes the developing fetus and is connected to the wall of the uterus. It is responsible for taking up nutrients, eliminating waste, and exchanging gas (e.g., oxygen) through the mother's blood supply. Recent research examined the placentas from births of several children and found that children who later developed ASD had a much higher rate (three times higher) of abnormal folds in their placentas (folds called trophoblast inclusions) than did children who did not develop the disorder (Walker et al., 2013). This is a "correlation" but a time line is established (folds before a later diagnosis), so this qualifies as a risk factor.

The finding conveys how risk factor or correlational research more generally can be very important and spawn additional research. In this example, we want to know:

- What theory or conceptual view might explain how these placental abnormalities are related to autism? That is, what processes are involved and how do these end up affecting the brain?
- What are the origins of these folds, leaving aside autism? That is, what places a placenta at risk, so to speak, for developing these folds (parent genetics, diet, hormonal abnormalities of the mother at a particular stage in pregnancy)?
- Do folds predict other health-related outcomes, even among those individuals not diagnosed later with autism?
- Can abnormal folds be identified in some other way (e.g., via the mother's blood) during pregnancy?
- Can placental abnormalities be prevented, and if they are does that change the physical or mental health outcome?
- What about exceptions. It is likely that some children from mothers with abnormalities of the placenta do not show autism or other diagnoses. Why?

There is more to the finding than the questions I have noted. A critical issue in developmental disorders such as ASD is early identification. Early identification can be used for early preventive intervention and also for more careful monitoring of someone to catch onset quickly if the disorder does come on.

Occasionally, novice researchers (but no one reading this text of course) are dismissive about correlations and the association of variables and say things like, "Oh, that's only a correlation?" The comment is well placed in one sense because it gives implicit admiration of true experiments, which seek causal relations and hence go well beyond correlation. The comment is wise too because it cautions about thinking that making a change in one variable will have any impact on the other, i.e., a correlation is not a cause. Yet, the comment also may be ill informed by using the word "only." Correlations and associations can be hugely important, and most of our understanding of physical and mental disorders begins with correlations. The associated features are eventually elaborated to develop a picture of what the full disorder looks like and what features might in fact be risk factors or causes. Worth mentioning in passing too is that much and sometimes most of what we know in natural, biological, and other social sciences are correlation, in part because we cannot manipulate many of the variables (e.g., in economics, meteorology, seismology [study of earthquakes and volcanoes]).

Risk factor is relatively easy to understand, although the term can put off individuals not trained in epidemiology and public health.[3] It is easy to remember that a risk factor is a correlation where the time line (what comes first) is clear. Thus smoking is a risk factor for later heart disease. And sometimes there are reciprocal relations where the time line goes "both" ways. So for example, depression (A) is a risk factor for heart disease (B) and the time line is clear. Once one has heart disease (B), now if he or she is also depressed (A), that is a risk factor for another heart attack and death. In other work A is a risk factor for B, but separately and interestingly B is also a risk factor for A. The nuances that the depression example suggests are not critical to the present discussion. Although if one has an idea that there is a reciprocal relation, that is a wonderful basis for investigation. The main point is that one source of research ideas is to identify factors that will predict later onset of a problem or predict an outcome of interest.

4.3.5: Protective Factor

More nuanced and less frequently studied is the notion of protective factor. Protective factor, as risk factor, has a time line feature so that it is some factor that is related to a later outcome.

> *Protective factor is a variable that prevents or reduces the likelihood of a deleterious outcome.*

Somehow for reasons that may initially be unknown, some characteristics within the individual, family, community, living situation, or other features prevent or decrease some outcome. The question raised by protective factor is obviously important. Once someone is born, or actually, before that point, what if anything can we do to decrease the likelihood that they will develop some disease/disability or fail in school?

> Protective factors can be conceived as the opposite of risk factors in the sense that they are negatively correlated with the onset of some later problem. So in any important sense, having a risk factor increases the likelihood of some outcome and having protective factor decreases the likelihood of the outcome.

The concept of protective factor often begins in the context of identifying special populations, namely, individuals who are at risk for a particular outcome. For example, children exposed to physical abuse and neglect are at risk for a variety of deleterious mental and physical health problems over the course of their lives; soldiers who have been exposed to combat are at increased risk of PTSD, and women who drink or smoke cigarettes over the course of their pregnancy are at increased risk for still birth, premature birth of the child, and birth defects.

What are those variables that separate those children, soldiers, and women at risk who do and who do not show the outcome?

What are the variables?

There might be some genetic factor, some environmental factor, and a variety of factors that reduce the likelihood of the outcome. Identifying these can be very important conceptually because it may hint at possible explanations or key mechanisms that could be involved. These factors are protective factors and begin with a group that is already identified as at risk. We investigate those who are at risk and show the anticipated outcome and those who are at risk who do not show the problems. The variables that are characteristic of this latter but not the former group are those that are called protective factors.

I mentioned methodology is a way of thinking, and risk and protective factors are two places where a lapse of that thinking is likely to emerge. Risk and protective factors are correlates. They establish something related to a later outcome. They are not causes. The lapse in thinking comes from protective factors in particular. Once one identifies a protective factor, programs are developed to build resilience and to protect people from an outcome. This well-intended goal makes an assumption that if one increased the protective factor, one would decrease some deleterious outcome. This latter statement could be true, but it requires research to study whether the protective factor bears any causal role. Identifying a protective factor is a critical step but is not enough.

For example, eating meals together as a family is associated with lower rates of risk behaviors among the teenagers in the family (e.g., lower rates of substance use, running away from home, violence) (e.g., Bisakha, 2010). In other words, family meals together might be a protective factor for many of the problems that can emerge in teen years. For example, in this particular study, two sentences appear as follows: "Family meals are negatively associated to certain problem behaviors. . . . Thus, programs that promote family meals are beneficial" (p. 187). Actually, the second sentence does not follow from the findings. It is a nonsequitur. There is no causal relation or implication by showing correlates, risk factors, or protective factors. In the case of this study, families that eat meals together, as opposed to those grabbing food on the run, eating individually, and not having many routines are very different for all sorts of reasons. Also, if one's child is never home or out buying and selling drugs, the number of family meals might be on the low end. That is, the problem (child behavior) may actually explain the so called protective factor (meals together) than the other way around.

In relation to the present discussion, one source of research is to identify factors related to the onset of dysfunction or some positive outcome (e.g., longevity, adaptive aging). Those factors might be risk factors (increase likelihood of some deleterious outcome) or protective factors (decrease likelihood of that outcome). This is an excellent focus of research as a way of testing or developing views of what influences might be operating and how. One has to be very clear about what one is studying (e.g., concurrent correlates, predictors, and causal agents) and what one is entitled to say once the demonstration is complete.

4.3.6: Causal Factors

The foci discussed to this point have been correlational. Correlation is not to be demeaned because findings are often provocative, intriguing, and spawn a great deal of research. For example, we know that prenatal hunger (from a time of famine when mothers were not well nourished) increased the risk of schizophrenia and depression in the offspring. "Only" a correlation (risk factor) but this raises scores of intriguing questions. Also, it is important to keep in mind as I mentioned that findings from most sciences (e.g., meteorology, cosmology, anthropology, epidemiology) are "correlational" because experiments where variables are manipulated cannot be easily done for many of the key questions.

Sometimes correlates (e.g., risk factors) when further studied can lead to our understanding of causal relations. In such situations, the variable moves from a risk factor to a causal factor. Of course, a variable does not "move" but it moves in how we classify the variable and what we can say about it. The variable was in the category "risk factor," for example, but might be able to go to the next level of understanding "cause."

As an example from medicine, high levels of cholesterol have been identified as a risk factor (correlational), but over time and with direct manipulation of cholesterol levels (in human and nonhuman animals) it was clear that cholesterol plays a causal role. Reducing cholesterol decreases the likelihood of heart disease. As an example from psychology, corporal punishment (moderate to severe) of one's child is a risk factor for later conduct problems, including psychiatric disorder. We know also that there is a causal role here too because several studies show that decreasing corporal punishment in the home decreases conduct problems (see Kazdin, 2005).

> For both heart disease and conduct problems, the story is more complex, but risk factor to causal relation is part of that story. Also, correlation/causal relations are often a matter of debate. For example, within the public, many see use of fossil fuel and climate change as correlated; most scientists see these as causally related—human impact has caused climate change.

Causal relations can refer to many different types of cause and causes that bear varied temporal relations to an outcome. For example, you are driving a car and go over a speed bump a little too quickly and a key part of your car (the engine) drops out on the road as you hear from the loud sound.

The site of a rather large pile of metal in your rear view mirror and your car feeling so much lighter suggest a causal relation between the event (engine loss) and the outcome (you coasting to a stop).

What is the cause? Well obviously if you had not hit the speed bump at 80 miles an hour, this would not have happened, so hitting the bump was one cause. Yet, the engine may not have been put in correctly to begin with (not fully tightened or not up to manufacturing specifications) or over time the connections may have become increasingly loose. Those are causes too. There may be a straw that broke the camel's back, but it was not that straw alone—but all the other straws and, well, the camel was pretty old and frail when piling on the straws even began so that influence is in the mix too. The answer hints at another point. Be careful in science and certainly in psychological science of the question, "What is *the* cause?" It is a lovely trick question that pivots on the word "the." Often there is no "the" cause but multiple causes. One might speak of one of the causes. But if some asks what is the *cause of schizophrenia*, or autism, or love of methodology, be careful; it can be a trick question. Some things seem to have a single, linear, and recognizable cause (e.g., rabies, broken leg, PTSD) but even here other influences may contribute to whether the obvious cause (e.g., bite from an infected animal, sports or car accident, exposure to sexual assault) leads to the outcome.

4.3.7: Key Criteria for Inferring a Causal Relation

Several criteria serve as guidelines for inferring a causal relation.[4] Table 4.3 provides key criteria that scientists use to infer a causal relation. They are viewed as guidelines and not as rigid requirements, and all of the requirements do not necessarily need to be met to infer cause (Ward, 2009).

> The most familiar criterion in laboratory and clinical settings experimental research is showing that a phenomenon can be altered by controlling an influence. For example, a great deal of research focuses on interventions (treatment, prevention, and educational programs) to reduce clinical dysfunction, to prevent the onset of dysfunction, and to promote learning and adaptive functioning. These studies focus on causal relations, i.e., making a change at the level of the individual, school, or community, for example, will lead to change in the outcome(s) of interest.

Table 4.3: Criteria for Inferring a Causal Relation between Variables

Criteria	Description
Strong association	Demonstration of a strong association between the independent variable or intervention and the dependent variable or outcome.
Consistency	Replication of observed result across studies, samples and conditions. Inconsistency might result from operation of a moderator and not controvert interpretation of critical construct. Consistency across studies facilitates drawing causal inferences.
Specificity	Demonstration of the specificity of the association among the intervention, proposed mediator, and outcome. Ideally, many plausible constructs do not account for the outcome, with the exception of one, which strengthens the argument that the proposed construct mediates change.
Time line	Demonstrating a time line or ordering of the proposed cause and outcome. The ordering and direction of influence must be clear.
Gradient	Showing a gradient in which stronger doses or greater activation of the independent variable is associated with greater change in the outcome. This is often referred to as a dose–response relation. If there is no dose–response relation (e.g., a qualitative or on-off effect rather than a gradient of effect), that does not refute a causal interpretation. The relation may be nonlinear and appear as "no relation" if only tested with a linear relation. Yet, where there is a gradient, this contributes to the ability to draw a causal inference.
Plausibility or coherence	A plausible, coherent, and reasonable process that explains precisely what the construct does and how it works to lead to the outcome. The steps along the way (from construct to change) can be tested directly.
Experiment	A causal relation is evident when one alters the independent variable (deliver it, vary key components to influence its effectiveness) and sees a change in the outcome. Intervening to change the variable is a strong way to test cause.
Analogy	Are there similar causal relations in other areas? For example, antibiotics for individuals with a strep infection often alter not only that infection but also psychological problems such as tics and obsessive compulsive disorder in those same patients (e.g., Murphy & Pichichero, 2002). Evidence that there is a causal relation is bolstered if there are analogous findings in related areas (and there are in this case).

These criteria are well known within science and emerged from a seminal paper (Hill, 1965) that has continued to serve as a source of discussion and debate (e.g., Höfler, 2005).

Intervention research focuses on causes of change, which may be different from, and not necessarily related to, the original causes that led to the development of the problem. For example, psychotherapy, surgery, and medication (e.g., aspirin) can "cause" change and eliminate a problem (e.g., anxiety, cancer, and headaches, respectively), although of course the absence of psychotherapy, surgery, or medication was not the cause of the dysfunctions to which they were applied. Related, in referring to a causal relation, it is important to bear in mind that there may be many causes. For example, to say that *a* causal relation has been shown between smoking and lung cancer is not the same thing as saying that smoking is *the* cause of lung cancer or the only cause. There may be many causes of lung cancer, and smoking is one of them—and a strong one at that. Yet, many people who have lung cancer have never smoked cigarettes, and many people who smoked cigarettes do not contract lung cancer.

We may know how to produce change (cause) even if we are not sure of the mechanisms involved. So when a randomized controlled trial shows that some intervention led to change, precisely what facet of the intervention produced change, or what intervening steps (e.g., affect, cognition, and behavior) led to the change in the target domain may not be clear. This was discussed in the context of construct validity, i.e., knowing that the intervention was responsible for the outcome but not knowing what aspect of the intervention was responsible. The study of mechanisms is coming next.

The guidelines to infer causality identified in Table 4.3 are not academic or relegated to the time (1960s) when the criteria were formulated. For example, I have mentioned that depression is a risk factor for heart disease. Many efforts have been made to establish the link and to show a causal connection.

- Does improving depression reduce heart disease, attack, or death from heart attack? Apparently no, not very much, or not consistently.
- What is the role of depression in heart disease?

A recent application of the criteria in Table 4.3 concluded that the link does not meet the criteria in a consistent way and cannot be considered causal (Meijer, Zuidersma, & de Jonge, 2013). The relation still needs to be elaborated, and the authors propose that the connection is related to factors that control cardiac risk, slightly beyond our present focus. Yet, the key point is that the criteria for causality were and remain a useful guide. Also, as a guide to research one might peruse the criteria for causality and use one of them to challenge or test a currently held view that causality is involved in some relation of interest or active study.

4.3.8: General Comments

Correlation, risk factor, protective factor, and causal relations are key concepts that guide research. I mention them here because they are a source of ideas. As one begins one's research, perhaps there is a topic of interest (e.g., obsessive behavior, heroism, friendliness, emotion regulation). One way to proceed further is to ask, what facet (e.g., correlate, risk factor) of that do I want to study? The concepts we have discussed in the previous section are one way to

consider how to proceed to the next step. The concepts are important to understand for interpreting research, but they also can be useful as a guide to developing the research idea. From the discussion, establishing causal relations is an ideal we seek and to achieve that manipulation of some phenomenon to show we can change it is the optimal strategy. Yet, it is always important to keep in mind manipulation of variables to demonstrate cause is not always possible or necessary.

4.4: Moderators, Mediators, and Mechanisms

4.4 Compare moderators, mediators, and mechanisms

Another source of research ideas is to focus on moderators, mediators, and mechanisms. These are worth delineating separately because of their importance, relation, frequent confusion, and rich sources of opportunities for developing studies. Table 4.4 provides an easy reference to summarize the definitions, and each is elaborated here. Each is a source of research ideas and an even better source of confusion.

Table 4.4: Moderators, Mediators, and Mechanisms Defined

Concept	Definition
Moderator	A characteristic that influences the direction or magnitude of the relationship between an independent and a dependent variable. If the relationship between variables x and y is different for males and females, sex is a moderator of the relation. Moderators are related to mediators and mechanisms because they suggest that different processes might be involved (e.g., for males or females).
Mediator	An intervening variable that may account (statistically) for the relationship between the independent and dependent variables. Something that mediates change may not necessarily explain the processes of how change came about. Also, the mediator could be a proxy for one or more other variables or be a general construct that is not necessarily intended to explain the mechanisms of change. A mediator may be a guide that points to possible mechanisms but is not necessarily a mechanism.
Mechanism	The basis for the effect (i.e., the processes or events that are responsible for the change; the reasons why change occurred or how change came about).
Moderated Mediation	The strength or direction of the relation of a mediator depends on some other variable. That other variable is a moderator.

NOTE: If these are difficult to remember, use the alternative definitions from everyday life. A moderator is someone who is the host on a quiz show; a mediator is someone who helps handle disputes (e.g., among divorcing partners, unions, and management).

4.4.1: Moderators

> Moderator refers to some characteristic that influences the direction or magnitude of the relation between the intervention and outcome.

If the effect of an experimental manipulation varies as a function of characteristics of the sample (e.g., sex, ethnicity, temperament, genetics, and neural activity) or setting (laboratory, clinic, at schools), these characteristics are moderators. We discussed moderators in the context of external validity or generality of findings. Will a particular finding generalize to all subjects, all ages, all settings, and all other conditions, or will it be moderated (influenced) by some other variable?

> We know about moderators from everyday life. For example, we know that all people who smoke cigarettes or gorge on high junk food daily do not suffer the likely consequences (e.g., cancer, heart disease, and may other untoward health consequences). That statement is an informal way of referring to moderators. That is, the relation of smoking or junk foods and the deleterious outcome is influenced by some other variable(s). Those other variables are called moderators.

Consider an example of a moderator pertinent to clinical dysfunction and everyday life. We know from our daily lives that a variety of annoyances and stressors can occur with some days much worse than others. These experiences on a daily basis can affect our mood and make our mood more negative. Makes sense—bad things happen and we are in a bad mood. Yet the relation is moderated by whether we are prone to rumination. Rumination is a way of responding to distress and involves focusing one's attention on one's negative emotional state and repetitively thinking about current feelings, causes, and potential consequences of that state (see Nolen-Hoeksema, Wisco, & Lyubomirksy, 2008).

In one study, college students kept daily diaries about unpleasant events (e.g., related to social or academic stressors) that occurred, their mood, and their ruminations. Considering days with more unpleasant events, negative mood was much more likely when rumination levels were also high (Genet & Siemer, 2012). That is, the level of rumination moderated (influenced, altered) the relation of unpleasant events and negative mood. This is important because we know now that mood is not merely a function of the negative events. Yes, they contribute, but one's style of processing those events is critical as well.

Moderators can be important in treatment studies. In fact, the dominant question that has guided psychotherapy research has been all about moderators, as illustrated by, "What treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances?" (Paul, 1967, p. 111). The question continues to receive prominence as the treatment agenda to guide research (e.g., DeRubeis et al., 2014; Kraemer, Frank, &

Kupfer, 2006; Roth & Fonagy, 2005). And one can see why the question of moderation is so important. Invariably no matter what form of treatment (e.g., psychological, pharmacological, surgical), some individuals do not respond.

- Who are these individuals?
- What is the moderator(s) that differentiates those who do or who do not respond or influences the degree of responsiveness would be a moderator?

4.4.2: Moderator Research

What is new about moderator research in relation to clinical phenomena is the range of moderators studied and how they are studied (e.g., more neuroimaging). For example, cognitive behavior therapy is an effective intervention for anxiety and depression.

A recent study found that treatment outcome was influenced by two moderators: severity of anxiety before treatment and how patients processed emotional facial expression, as evaluated during a brain imaging (fMRI) task (Doehrmann et al., 2013). Precisely why and how emotional processing moderated the outcome is not known but may generate important leads about social anxiety and perhaps as well as how treatment achieves its change and how the brain is altered by treatment.

Perhaps a relevant answer to some of the questions is another moderator study showing that facets of brain functioning (glucose metabolism in the right anterior insula) in areas of the brain that relate to depressive symptoms as well as affective and cognitive processes (e.g., emotion regulation, decision making, cognitive tasks) can serve as biomarkers, i.e., biological moderators of treatment (McGrath et al., 2013). In this study, depressed patients were evaluated (using positron emission tomography or PET scan) and assigned to treatment (medication or cognitive therapy). The level of glucose activity (over or under) predicted responsiveness to treatments. For example, cognitive therapy patients with low activity responded much less well to treatment and were more likely to experience remission (return of the dysfunction). This is enormously important as a line of work because directing patients to treatments likely to work and away from those unlikely to work are pivotal goals of moderator research. In addition, brain activity biomarkers may suggest why and how moderators may work and targets for intervention.

The importance of moderators is easily conveyed in more everyday life examples that show the relation between variables can be drastically altered based on a moderator. We know, for example, that a diet rich in fish can lower the risk of heart attack. Dark fish (e.g., red salmon compared to cod or sole) especially is beneficial because of the high content of omega-3 fatty acids "good fats." The benefits of fish, dark or not, appear to operate by lowering inflammation (related to heart disease), blood pressure, and cell damage. So fish is healthful, generally speaking! There is a huge moderator I have not mentioned that changes the story. The benefits of fish are moderated by how the fish is prepared. If the fish that is eaten is broiled or baked, the risk of heart attack is reduced. If the fish is fried, the risk of heart attack is increased (Belin et al., 2011). This was evident in a prospective study of over 84,000 women (ages 50–79) followed for an average of 10 years. In short, whether fish help or hurt (decrease or increase risk of heart attack) depends on how it is prepared. Clearly moderators can be very important.

Moderators are used as a basis for designing studies. Consider what variable might make a difference or change the relation between two other variables? Or you read a finding and say, "that cannot be true or always true." Now go ahead and show when the finding does and does not hold, i.e., because of some other third variable or moderator.

4.4.3: Mediators and Mechanisms

I have mentioned *cause* or *causal relation* and that is a useful point of departure for describing mediators and mechanisms. We begin with an intervention that we know causes some change. For example, exercise is an effective intervention to reduce clinical depression as attested to in controlled trials (e.g., Blumenthal et al., 2007). From such demonstrations, we can say that an intervention caused the change, as that term is used in science. Demonstrating a cause does not say why the intervention led to change or how the change came about. To evaluate how change comes about, research often looks at mediators and mechanisms.

Mediator is a construct that shows a statistical relation between an intervention and outcome.

This is an intervening construct that suggests processes about why change occurs or on which change depends. Mediation is evident when several conditions are met:

1. The intervention (e.g., exercise) leads to change on outcome measures (e.g., depression).
2. The intervention alters the proposed mediator (e.g., perhaps stress level is proposed to serve that role).
3. The mediator is related to outcome (stress level is related to symptoms).
4. Outcome effects (changes in depression) are not evident or substantially less evident if the proposed mediator (stress in this example) did not change.

It is possible that exercise only was effective if stress level changed in the process of treatment. That would suggest mediation. It is also possible that exercise helped reduce depression even if stress levels did not change, which would suggest that something else about exercise may account for the change.

These relations convey that change was mediated (e.g., correlated with, depended on) by some construct. Figure 4.1 shows a simple schematic to conceptualize mediation and puts into a picture the thousand words I used to describe the same material.

Figure 4.1: Mediation Model

Mediation model involving independent variable, dependent variable, and a proposed mediator.

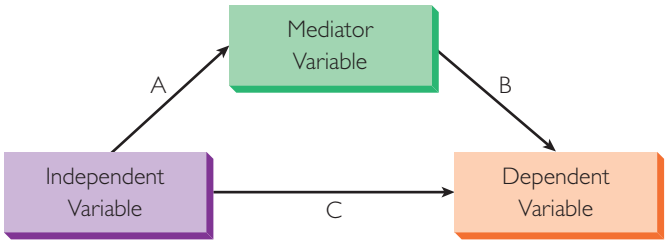


Figure 4.1. Simple illustration of a mediation model and a hypothetical example involving an independent variable or manipulation (exercise) and dependent variable (depression) and a proposed mediator (stress reduction). The hypothesis is that the independent variable or manipulation (e.g., exercise) is effective in changing the dependent variable (e.g., clinical depression). That can be tested by a randomized controlled trial that evaluates exercise and a control or comparison group (e.g., another treatment, no treatment). The study of mediation goes further and hypothesizes that exercise works or achieves its effects because of some intervening process (e.g., stress reduction). That is, exercise works because it reduces stress and that is why depression is decreased. This means further that exercise may not work at all or very well unless stress is reduced. This is a hypothetical example. Many statistical tests are available to evaluate the relations among variables (A, B, C in the figure) to see if the results are consistent with the hypothesized mediator as an explanation the connection (C) between the intervention and outcome.

Even when the conditions are met, considerable ambiguity can remain about the precise role of the mediator. Mediation may be partial (some relation but not very strong or complete). Also, the mediator might serve as a proxy (stand for) for one or more other variables with which it is correlated. More critical, the mediator may not and usually is not intended to explain precisely how the change comes about. Once a mediator is identified, the investigator may speculate what about that mediator leads to change and how that change comes about, but the demonstration of a mediator per se usually does not show that latter level of detail. Thus, mediation gets us closer to understanding what might be involved in change processes. In the hypothetical example, it could be changes in stress. That can be very helpful in directing next steps.

By and large, mediator is a statistical relation and points to key constructs that might well explain processes involved in change.

> Mechanism refers to a greater level of specificity than mediator and reflects the steps or processes through which the intervention (or some independent variable) actually unfolds and produces the change.

Mechanism explains more about underlying processes and how they lead change. Multiple studies may be essential to find out how change occurs and these studies often combine basic (e.g., animal laboratory) as well as clinical studies.

For example, we know that antidepressant medication influences a special protein that stimulates growth and differentiation of neurons and synapses in the brain, especially the hippocampus (see Duman & Aghajanian, 2012).[5] We also know from several studies with humans that major depression is characterized by low levels of this protein (BDNF) and that these levels increase after successful antidepressant treatment (Sen, Duman, & Sanacora, 2008). In nonhuman animal studies (rodents), antidepressant effects can be manipulated experimentally to isolate the processes involved in change (e.g., by placing the protein into the hippocampus directly, by gene-knockout studies and blocking studies negating the operation of the protein). (Depressive behavior in animals is often evaluated by learned helplessness responses and forced swimming tasks.)

Also, in keeping with our ongoing example of exercise, manipulation of exercise in animal studies leads a therapeutic-like antidepressant effect (task performance) and alters the protein considered to underlie depression and change (e.g., Duman, Schlesinger, Russell, & Duman, 2008; Shirayama, Andrew, Chen, Russell, & Duman, 2002). These studies move very far in identifying precisely what is involved in successful intervention and symptom change. Exercise affects a protein associated with depression and may be one way in which exercise operates in human depression. This is more specific than mediation (statistical relation of constructs) and begins to point to underlying processes that are altered. More work is needed of course. How does a change in a specific protein lead to changes in affect, behavior, and cognitions associated with depression?

4.4.4: Tutti: Bringing Moderators, Mediators, and Mechanisms Together

For presentation, I have treated the concepts separately and simply but they go together. As a case in point, moderators can help elaborate mediators and mechanisms of action. Consider an example of the effect of experience during childhood on subsequent criminal behavior, where a genetic characteristic is a moderator. Children with a history of physical abuse are at elevated risk for later

antisocial behavior (e.g., criminal acts, aggression, domestic violence) (Child Welfare Information Gateway, 2006), even though most people who are abused as children do not engage in antisocial behavior later in life. A genetic characteristic moderates the relationship. Maltreated children with a genetic polymorphism (related to the metabolism of serotonin) have much higher rates of antisocial behaviors than those without this polymorphism (see Anholt & Mackay, 2012).[6] My description so far makes it easy, namely, one moderator. Yet more findings can be brought to bear. First, the relation I have noted (the polymorphism is a moderator) applies to boys rather than girls. Also, when maltreatment is severe the moderator has less of an influence, i.e., severe maltreatment promotes antisocial behavior with or without the polymorphism. One can see why one needs more research and finer-grained analyses. Findings are not merely of the nature boys are different from girls.

So far, this is a fascinating illustration of moderation—the outcome physical abuse depends on another variable (moderator). However, closer scrutiny may hint at mechanism. The gene that encodes a serotonin related enzyme (the MAO-A enzyme) is linked with maltreatment victimization and aggressive behavior (Caspi et al., 2002). A rare mutation causing a null allele (absence of the critical characteristic) at the MAO-A locus in human males is associated with increased aggression. Gene knockout studies in nonhuman animals show that deleting this gene increases aggression. Restoring this gene expression decreases aggression.

In one sense we have identified a moderator: The influence of an independent variable (abuse in the home) and outcome (antisocial behavior years later) is moderated by some other characteristic or variable (MAO-A allele). Clearly, we have much more because the moderator points to possible genetic, neurological, and molecular underpinnings (see Anholt & Mackay, 2012; Heinz, Beck, Meyer-Lindenberg, Sterzer, & Heinz, 2011). The ability to increase or decrease aggression with genetic manipulation meets many of the criteria for causal relation noted earlier (Table 4.3). We do not know how the allele and abuse traverse specific steps through which aggression emerges, but we are getting closer by showing that manipulation can lead to change. Also, other findings show the neural mechanisms through which the genetic influence is likely to operate (Meyer-Lindenberg et al., 2006). The MAO-A allele is associated with diminished brain circuitry related to impulse control that would promote aggression. In short, this example illustrates how the study of a moderator might well lead to insights about mediation and then to possible mechanisms of action.

It is possible that the mediator or mechanism of change varies as a function of a moderator variable, a phenomenon referred to as *moderated mediation* (Muller, Judd, & Yzerbyt, 2005; Preacher, Rucker, & Hayes, 2007). (If at any point, the reader is confused about moderation and mediation and how the terms are used, there is a help center available 24/7 online for those in need [Jose, 2008].)

> *Moderated mediation occurs when the strength (or direction) of the relation of the mediator to outcome depends on the level of some other variable.*

Understanding this begins with recognition that a given outcome can be reached through different means (mediators). For example, from well-controlled experiments, we know that intelligence quotients (IQs) can be increased in children in diverse ways, including dietary supplements, early educational interventions, interactive reading with a young child, and sending a child to preschool (Protzko, Aronson, & Blair, 2013). Thus, a single outcome (higher IQ) has many paths, and these paths may reflect different mechanisms leading to an outcome. The different mechanisms depend on other variables (moderators), which are experiences to which the children are exposed.

Moderated mediation is evident when subgroups are identified or emerge. This was evident in a psychotherapy study (12 months of psychodynamic therapy) in which treatment was evaluated with measures of symptom change as well as brain metabolism (Lehto et al., 2008). Atypical depressed patients, categorized in advance, showed metabolic changes in response to treatment, but other depressed patients did not. This can be discussed as an example of different mediated processes in the brain as a function of subtype of depression.

4.4.5: General Comments

Many concepts and examples were provided; each can serve as a way of prompting the focus of a study. In developing a study, consider what may influence the findings so that they are stronger or weaker based on some other variable (moderator). It is likely that many influences we consider to be universal in fact vary as a function of some other variable, including such strong influences as culture and ethnicity. We discussed this previously by noting that findings obtained with WEIRD college students (Western, Educated, Industrialized, Rich, and from Democratic Cultures) often do not generalize to individuals in other settings and of diverse cultures (e.g., Henrich, Heine, & Norenzayan, 2010a, b). That is, culture (and other characteristics) can be a moderator. Now mediators might be proposed to identify what precisely about culture might be a critical construct to help explain the processes involved.

Moderation is pervasive, which is why methodologists are fond of the expression, "Everything in moderation." In stating this, they do not mean Aristotle's advice of not doing anything in excess. The methodological version is more profound. As you search for research ideas, look to

findings of interest and ponder what might moderate the relations that were demonstrated. Be critical of your favorite study (especially if it is your own) and challenge the finding by noting for whom that finding is not likely to occur and why. Now you have the idea for a study.

Mediation too receives considerable attention as the impetus for a study. Here one identifies possible explanations of the effects that have been demonstrated. This is the substantive (rather than methodological) part of construct validity.

Precisely what is the construct that explains the effect? Could it be expectations, novelty, reduction of stress, increase in hopefulness, and so on? Mechanism of action or precisely what underlying processes are involved is yet a more fine-grained analysis and relatively few studies work at that level.

Among the key issues to remember is that it is likely there is no one moderator, mediator, or mechanism to explain a given relation. In all of the research highlighted here, it is important to have a theory or hypothesis. One does not blindly throw in moderators or mediators just to see. One begins with a view about what is going on and why, and moderator and mediator studies test that.

4.5: Translating Findings from Research to Practice

4.5 Identify characteristics of the full process of translational research

The notion of translational research or ways of moving basic findings so they reach people in need has received increased emphasis in recent years. This is worth some discussion because one source of ideas for a study maybe "translational." That is, consider how a finding might be extended to clinical use or even larger-scale application. The full process of translational research is characterized as moving a research finding from "bench" (basic, laboratory research, often with nonhuman animals), to bedside (application for patients), or to the community (application on a large scale, perhaps a population scale if relevant to public health). Before elaborating these concepts, consider some background briefly.

4.5.1: Basic and Applied Research

Long before the term, translational, there has been a well-recognized distinction between basic and applied research.

Basic research usually has one or more of the following characteristics:

- Provides a test of a proof of concept or theory to identify what can happen
- Makes an effort to understand a phenomenon of interest under highly controlled conditions
- Isolates processes or variables in ways that might not be how they appear in nature
- Uses nonhuman animal models that allow special evaluation or observation of a key process
- Uses special circumstances (e.g., procedures, equipment) that allow control or assessment of effects not otherwise available

For example, basic research studies on mice might control the experiences they have being raised under varied mothering conditions or whether some genetic component is "knocked out" to see whether it influences later aggression, obesity, or cooperation. Also, many such studies are a proof of concept test. The goal is to identify what can happen. The terms "bench research" and "lab research" also have been used to characterize basic research, and there have been no formal or consistent delineation or distinctions among these terms and basic research.

As an illustration, we have known from extensive years of basic research with many species that calorie restricted diet, can slow the aging process, and reduce rates of death from many of the diseases associated with aging (e.g., Heilbronn & Ravussin, 2003; Roth & Polotsky, 2012). We refer to this as basic research because the goal was to provide artificial circumstances, quite different from what they would be in nature, and evaluate how rather severe restriction influences aging. This work has been important not merely to show that aging can be slowed, but also to understand the biological underpinnings. Thus, research has looked at precisely what calorie restriction does at cellular and molecular level to identify the mechanisms involved (e.g., Kume et al., 2010).

That basic research is fundamental but does not instantaneously lead to findings that help us right now to age more slowly. The calorie-restricted diet (20–40% reduction in calories) is not readily feasible because it is much more than merely cutting back on breakfast nachos and snacks while watching movies. Yet, that is not the criterion for evaluating the value of the research. Rather the goal is to understand aging, and this can be accomplished by describing and explaining what happens with calorie-restricted diet. Perhaps once we understand precisely how calorie restriction works to alter aging, we might be able to influence or control antiaging effects without the calorie-restricted diet. That is, calorie restriction does many things to the body and perhaps those can be achieved in other ways, i.e., without calorie restriction.

4.5.2: Distinguishing Applied Research from Basic Research

Applied research is distinguished from basic research and usually has one or more of these characteristics:

- Provides a test that focuses on an applied problem that may be of direct benefit to individuals

- Tests what can happen in contexts and real-life settings (e.g., schools, clinics, at home)
- Makes an effort to have impact (e.g., reduce symptoms, improve test performance or competence) and may have a practical component of helping in addition to answering an important research question
- May isolate influences (e.g., components of a prevention program) but also looks at intervention packages (e.g., cognitive behavior therapy) that are complex interventions with many components to see if there is overall impact
- Is concerned from the outset of generality to everyday settings

As was the case with basic research, not all characteristics are required in any individual study. Also, once having highlighted characteristics of basic and applied research, it is clear that the distinction is clear at the extremes. For example, it is easy to call research basic when memory and learning are studied in some nonhuman animal model on a special task and to call research as applied when memory and learning are the subject of a large-scale intervention design to improve student performance in math. Once one leaves the margins, the distinction is blurry. The blurriness is "good" in relation to the chapter because it conveys a bipolar continuum where basic (e.g., on the left) and applied research (on the right) can vary and along multiple dimensions (e.g., how realistic the setting is, how much like the setting to which one might want to generalize). This continuum provides many opportunities to do research.

In clinical psychology, there has been a long-standing distinction between basic and applied research in the context of psychotherapy. The most recent incarnation has used the terms "efficacy" and "effectiveness research." *Efficacy research* indicates that a treatment is conducted under highly controlled conditions, often with nonclinical samples. Clients are screened to maximize homogeneity of the subjects and to provide a strong experimental test. The emphasis is on internal validity and all the ways to exert control to demonstrate an effect. *Effectiveness research* is designed to evaluate treatment in clinical settings, with "real" patients, and under conditions more routinely seen in clinical practice. While internal validity still is important in such studies, they begin with a strong interest in external validity, i.e., developing interventions that can be applied in every day settings.

> One can see right away that this is a bipolar continuum because research can vary in the extent to which it leads toward one side (efficacy) or the other (effectiveness) in varying degrees.

Also, there are multiple characteristics of a treatment study (who serves as clients, as therapists; degree to which therapy mimics how it might be used in clinical practice, whether clients are paid for participating treatment as often the case in funded research or are charged for treatment, and so on). Each characteristic may move toward highly controlled and artificial (compared to "real-world" applications) or closely follow how the intervention would be used in clinical practice.

Although efficacy and effectiveness are heavily discussed in clinical psychology, the issue characterizes research that has applications in many different contexts. For example, educational interventions including what can be done to improve learning, school performance, and graduation rates at all levels of schooling are the foci of many well-controlled experimental studies (efficacy). Here too the questions and challenges include whether the results can be extended to school settings under the conditions without such controls, monitoring, and care (e.g., Kratochwill & Shernoff, 2004; Kremer, Brannen, & Glennerster, 2013).

4.5.3: Translational Research

A concern with basic research has been that many findings take a long time, often decades, to move from the lab to helping people in everyday life. This applies to many areas (e.g., psychological, medical, and educational interventions). Translational research emerged in an effort to move findings from the lab to clinics more systematically and quickly. This is discussed in medicine in which basic biological research may not get translated very quickly to medical applications. That is the context for referring to translational research *"bench to bedside,"* where "bench" is equivalent to "laboratory" or "basic" and bedside is equivalent to "directly applied."

There is no single agreed-upon definition of translational research, and several have been provided (e.g., Bardo & Pentz, 2012; Woolf, 2008). It is not so much that the different definitions disagree but rather that many different kinds of research qualify and hence the type that is emphasized can vary. It is better to consider the key characteristics of transitional research than any single definition.

> Translational research encompasses basic and applied research issues but has some new features too. The effort is to unite understanding processes (e.g., clinical dysfunction, disease) and moving them to therapeutic interventions. That is, from the outset a goal is to develop collaborations that have in mind both basic research and its extension.

For example the National Institute of Health has a "Bench to Bedside" (B2B) research program (see http://www.cc.nih.gov/ccc/btb/). The goal is to foster collaborations or teams of researchers to work together so that the gap between basic findings and their extension to clinical care can move more systematically and quickly than the normal process. That normal process is one in which basic researchers

and applied researchers have little contact and that is part of the problem as to why research findings do not usually get translated very well. The collaborations are designed to address barriers, "such as the traditional silos between basic and clinical researchers" (Web site above, no page).

Translational research also is novel in moving research in both directions so that it is not only from basic research to application, but from application to basic studies.

- What can we learn from clinical work, from existing databases, or from complementary and alternative medicine (e.g., diet, micronutrients) that may be effective?
- From what we observe in clinical settings or practice (bedside), what can we scrutinize better in basic research to understand what may be going on?

For example, we know that exercise has all sorts of mental health benefits, including treatment of a variety of psychiatric disorders (e.g., Wolff et al., 2011). We can go back to the lab and try to understand precisely how this works (e.g., animal models, brain studies, genetic moderators). Perhaps our basic understanding of exercise can improve on exercise but also identify processes that might be affected or altered other ways as well (e.g., diet, medication). As one illustration, we know that exercise alters some neurotransmitters implicated in many psychiatric disorders (e.g., Wipfli, Landers, Nagoshi, & Ringenbach, 2011), but scores of other biological markers also change with exercise (e.g., inflammation, blood cells, circulating proteins in the blood) (e.g., Shanely et al., 2013). Much more laboratory work is needed to identify how exercise influences psychological dysfunction. But the larger point is the one to emphasize, namely, translational research includes bedside to bench as well as bench to bedside studies. The thrust of translational research includes keeping these sides closely connected.

4.5.4: Further Consideration Regarding Translational Research

Although "bench to beside" is the key phrase that characterizes translational research, the additional term conveys a broader thrust. Translational research includes "bedside to community" which means bringing the findings and applications to others on a larger scale. This means taking bedside findings, i.e., research that can help individual patients or groups of patients in relatively small studies to the level of the community. Community here refers to public health interventions that can be scaled up. There are many models for this. Vaccinations may be among the most familiar in which very basic studies are done (e.g., many animal studies, evaluations of underlying processes), and these move to small scale or isolated applications to monitor their effects. Eventually these move to community-wide applications. This is a very slow process. Translational research is designed to speed this up by structuring the processes (e.g., via collaborations) and specifying the need for fluid and bidirectional boundaries (back and forth from basic to applied and back again).[7]

Translational research is critical in clinical psychology and related areas of application (counseling, education, psychiatry). The development of evidence-based treatments illustrates the problem and efforts toward solutions. There are now many evidence based psychotherapies for children, adolescents, and adults (e.g., Nathan & Gorman, 2015; Weisz & Kazdin, 2010). These are well-researched treatments often with very elegant controls, meticulous analyses, and clear demonstrations. In some cases, but certainly not all, the highly controlled studies are more toward the "bench" side of research (research settings). All sorts of questions have emerged about moving these to the "bedside" (patient care in everyday clinical settings). It is still the case that clinical practice is not using the most well-studied treatments, and when the techniques are used in clinical practice, their effects are often diluted. So we can see we have a bench to bedside issue, i.e., extending controlled research findings to clinical settings. Much research has turned to dissemination of evidence-based treatments, which includes training practitioners to use treatments and evaluating treatments in clinical settings (e.g., highly select patients without multiple disorders), and restructuring treatments themselves (e.g., into modules) that are more bedside friendly (Weisz, Ng, & Bearman, 2014).

Dissemination is the "bench to bedside" part, namely, getting well-studied treatments so that they are used in clinical practice. Yet we also have an even greater bedside to community issue. Most people in need of psychological services receive no treatment at all, not even those nonevidence-based treatments that are in such widespread use. We must do much more to get our treatments so that they are not only effective, but are used in clinical applications, and on a larger scale (Kazdin & Rabbitt, 2013). Large-scale and community-level findings have been more the domain of public health and social policy than clinical and related areas of psychology (e.g., counseling, school, educational). Yet, the boundaries (of disciplines) and type of work in which they engage (bench, bedside, community) are blurred.

Psychological research is more likely to focus on the back and forth of research from bedside and bench. For example, we do not really understand why psychological treatments work. We need basic studies to reveal mechanisms and to do this research has often drawn on animal models to study interventions for anxiety and depression (Davis, Myers, Chhatwal, & Ressler, 2006; Duman & Aghajanian, 2012). Such work has already led to improvements in treatment research with humans.

Translational research is discussed in the present chapter because it is very much related to the source of ideas. To begin, one source of ideas is to consider findings from basic research (e.g., on learning, memory, emotion, implicit attitudes) and how they might be applied to studies in ways that relate to everyday life. There may be intriguing experimental findings, and perhaps one can study them in more of an applied context. For example, priming studies in social psychology experiments set up artificial (basic, controlled) contexts to see if human behavior can be influenced and by ways outside of the awareness of the participants (Bargh, Schwader, Hailey, Dyer, & Boothby, 2012). This is extremely important work that provides a proof of concept test and evaluates fundamental information and brain processes.

Can the finding be moved closer to bedside? Perhaps the research could be extended to psychotherapy processes where some priming is used to improve some facet of the treatment process such as patient disclosure of information or the therapeutic relationship (e.g., Grecco, Robbins, Bartoli, & Wolff, 2013). In short, basic findings can be used as a basis for translating (applying) key principles to more applied contexts. Even more ambitious, can priming be used on a large scale for the public good (e.g., improving nutrition, reducing energy consumption)? Large-scale application moves from research into other areas, such as social policy and legislation.

In terms of sources of ideas for research, one might develop a study by moving from application (e.g., bedside) to basic research (e.g., bench). Identify applied findings one finds interesting (e.g., a particular intervention decreased suicidal ideation or unprotected sex; or reported use of emotion regulation strategies influenced their response to stressors in everyday life). Now ask a basic research type of question about "why" and perhaps begin with a little theory of what you believe is going on. Now design a "bench" or laboratory study (e.g., perhaps college students, MTurk) where conditions are dissected or controlled to permit a test of theory.

There is strong interest in translational research among funding agencies, researchers, the public at large, and policy makers.[8] Among the interest is the question from the public and policy perspective—*what are we getting from all the research we are funding, and are we helping people?* It is easy to answer the question with a strong "yes," but it is equally easy to identify enormous delays in moving evidence to application and to point to large swaths of people in the United States and throughout the world who are not receiving preventive and treatment interventions that we have known to be effective for some time. The comparison to see where new procedures reach the public more efficiently is evident in business where innovations (e.g., better smartphones and tablets, screens for TV viewing) get to the public as quickly as possible. As you do your research or read the research of others, consider where it might fall on the continuum of bench (lab), bedside (applied or clinic setting), and community (larger-scale application as in public health) and what might be next steps at the level (bench, bedside, community) that most interests you.

4.6: Theory as a Guide to Research

4.6 Define theory

The concepts such as correlate, risk factor, moderator, mediators, and others do not convey the full range of foci of investigations, but they illustrate overarching ways of identifying research problems and sources of ideas. More generally, the concepts show the important movement from merely establishing a relation to elaborating critical features about that relation. The progression of research from description to explanation, from correlation to cause, and from research to application as described to this point may inadvertently imply a crass empiricism, i.e., one merely tests different types of relations among variables to see what role they play, if any, or one takes a finding from one area (e.g., bench) and just tests it in some application (e.g., bedside). Underlying the concepts that guide research (and material in Table 4.2) is the investigator's theory and that is a critical part of the research process.

4.6.1: Definition and Scope

> Theory, broadly defined, refers to a conceptualization of the phenomenon of interest.

The conceptualization may encompass views about the nature, antecedents, causes, correlates, and consequences of a particular characteristic or aspect of functioning as well as how various constructs relate to each other. There are many related terms that seem to serve as theory or conceptual underpinnings of a phenomenon of interest. Prime examples are terms such as approach, conceptual view or model, theoretical framework, and working model. Theory and these other concepts are used with great variability; they also tend to be fuzzy and overlap. For present purposes and as a guide to developing research, theory is an explanation of what is going on, why and how variables are related, and what is happening to connect those variables in specific ways.

> Theories can vary in their scope of what they are trying to explain. In clinical and counseling psychology, theories of psychopathology and personality have been a central topic in which diverse facets of human functioning are explained.

Historically, psychoanalytic theory illustrated this well by posing a variety of constructs and mechanisms that

were designed to explain intrapsychic processes, child development, parent–child interaction, dreams, slips of the tong, and performance in everyday life, psychopathology, character traits, and more.

Research in psychology has moved away from broad, all-encompassing views. More circumscribed theoretical views characterize contemporary research in an effort to develop specific models or integrated sets of findings and relations among variables. The conceptual views focus narrowly on some facet of functioning rather than to develop a grand theory to explain so much. For example, the models may explain the relation between specific characteristics and a disorder (e.g., hopelessness and helplessness in relation to depression) and how these characteristics lead to other features of dysfunction.

A theory provides a tentative explanation of how variables are related. For example, mother depression and child adjustment and social functioning are related (Goodman et al., 2011). A theoretical statement may propose and test how and why these are related. It may be that the link is genetic in some simplistic way (the biological propensity in the parent is passed on in the infant), or biological in some other way (e.g., hormonal abnormalities perhaps induced by stress during pregnancy that had enduring effects of the functioning of parent and child), or child–parent interaction (e.g., poor bonding and attachment). These all may be important; a theory tries to explain not only what the connection might be but why. Here tests of mediators might well be applicable. A test of a mediator requires a little theory as to the connections between independent and dependent variables.

4.6.2: Theory and Focus

Theories can be broad too of course. Broader theories may be proposed that account for different types of disorders and how multiple variables come together and operate. One might include the interplay of biological (e.g., temperament), psychological (e.g., social relations), and contextual (e.g., living conditions) into a network or larger model that explains how depression or other disorders come about. There is interest in psychopathology in transdiagnostic models, i.e., explanations that go across different disorders or psychiatric diagnoses. Among the reasons is that disorders (e.g., depression, anxiety) often have overlapping symptoms and multiple disorders often are present in the same individuals (comorbidity). Also, there is now genetic evidence that indicates surprising commonalities among different disorders (e.g., Serretti & Fabbri, 2013). Broad and narrow theories may be needed to explain how similar beginnings can yield to dysfunction and then why these branch off into different dysfunctions or symptom patterns. Research on any facet of this could serve as a valuable source of ideas.

Apart from the scope of theory, the focus may vary. Consider three examples.

1. Theory may focus on the origins and nature of a clinical dysfunction or behavioral pattern. Here the theory would consider conceptual underpinnings and hypotheses about the likely factors leading to the clinical problem or pattern of functioning, the processes involved, and how these processes emerge or operate. Perhaps the theory would consider various risk and protective factors, paths and trajectories, and how early development results in subsequent dysfunction.
2. The theory might focus on factors that maintain a particular problem or pattern of behavior. Here the theory might consider the factors that might operate to influence, sustain, or shape the way in which the problem is continued, long after the onset is established. Perhaps the theory would focus on how, why, or when relapse occurs, i.e., why a pattern is not maintained.
3. The theory might focus on change as in therapeutic change or changes in development. In the context of therapy, the theory might consider the necessary, sufficient, and facilitative conditions on which change depends. There are many other areas where theory would be relevant. In each case of course, the reasons are proposed to explain how and why the relations of interest occur.

The notion of theory can be overwhelming. It implies broad conceptual frameworks of how the universe came into being. Also, we have in the back of our mind all-encompassing theories that required a special brilliance (e.g., theory of relativity, evolution). Broad theories can be valuable if they ultimately can be shown to make testable predictions about a phenomenon. Yet, small and narrow theories are very valuable. They are important in their own right and can be expanded as research allows. So in your study of moderators (or mediators), the "theory" part is your view of why something would make a difference. Theory is the opposite of saying, "just because"—we need a statement of what led you to think that and then what predictions are you testing based on that.

A way to practice what is required is to think of someone in your everyday life who engages in a behavior you find particularly enjoyable or annoying. Now ask yourself, "why do they do that?" Our answer is a mini-theory. That is the easy part. Now move to developing one or two ways that might test the theory.

For a theory to be a scientific theory, it must generate testable hypotheses or predictions. In this hypothetical example, the theory and prediction might be something like, "If the person does the behavior for the reason I am proposing, a good test of that would be to see if he or she does x or y in response to some other situation" or "what

would make me give up my explanation of why the person does that?" In everyday life, we usually keep our theories because they are not put to the test or because cognitive heuristics help us maintain them, even in the face of counter or conflicting evidence. In science, we devise the theories for the purpose of making predictions, testing them, and revising the theory as needed.

4.7: Why Theory Is Needed

4.7 Report the relevance and benefits of theory in research

A goal is to understand human functioning and to achieve that we do not merely accumulate facts or empirical findings. Rather, or in addition, we wish to relate these findings to each other and to other phenomena in a cohesive way. For example, an investigator may demonstrate that there are sex differences regarding a particular disorder, personality characteristic, or cognitive style. However, by itself sex differences are not necessarily interesting. A theoretical understanding would pose how this difference develops, what implications the difference may have for understanding biological or psychosocial development. For example, recent brain imaging research has identified differences between women and men in response to hearing infant cries (De Pisapia et al., 2013). Women, whether or not they are parents, are more likely to shift their attention in response to the cry; men continue in the state they were in (in response to control noises). With this finding, all sorts of questions about processes involved and the scope of the differences are raised. Knowing the specific changes in activation, one might theorize the scope of differences that might be evident beyond responding to infant cries. Theory can help here by suggesting what might be involved and how that would be manifest in other male and female differences. From the standpoint of research, theoretical explanations guide further studies and the data generated by the studies require emendations of the theory. This is an important exercise because theory moves us to implications beyond the confines of the specific empirical relations and the restricted conditions in which these relations may have been demonstrated.

One can be more specific about why theories are needed and the benefits that derive from them.

1. The first theory can bring order to areas where findings are diffused. For example, consider the area of psychotherapy. We know there are hundreds and hundreds of psychological treatment techniques and the number continues to grow (e.g., Kazdin, 2000). Theory could bring unity to this area. Perhaps there is a small set of common mechanisms or processes that could be identified that span several treatments. Assume for a moment that many or most of the treatments are effective (although the vast majority have never been studied in any empirical investigation), and it is unlikely that all the treatments work different reasons. Indeed, it is quite unparsimonious to begin with that thought. There might be a few theories that account for how change comes about and that unite the disparate treatments and their findings.
2. The second theory can explain the basis of change and unite diverse outcomes. Again, using therapy as an example, all sorts of changes occur in treatment. Of course, therapy changes various social, emotional, and behavioral problems for which individuals often seek treatment (e.g., depression, anxiety). In addition, therapy improves symptoms of physical health, including indices of serious disease (e.g., heart disease, diabetes) (e.g., Hardcastle, Taylor, Bailey, Harley, & Hagger, 2013; Harkness et al., 2010; O'Neil, Sanderson, Oldenburg, & Taylor, 2011). How can these effects occur? The answer entails a theoretical statement, which is merely a tentative explanation that can be tested. Such a statement when elaborated empirically could greatly improve our understanding of many facets of human functioning, beyond psychotherapy.
3. The third theory can direct our attention to which moderators to study. In any area, there are an infinite number of moderators that might be proposed. The standard litany would be sex, age, gender, socioeconomic class, and the list could continue to encompass all characteristics of people and the conditions and contexts in which they live. For example, marital satisfaction could be influenced by scores of characteristics of each partner (e.g., style of emotional regulation, attachment style developed in childhood, sibling relations, histories of their parents, current living conditions, personality of each person, education, similarity of each partner on any one of the above characteristics, and an endless so on). We do not want research merely to catalogue what factors do and do not serve as influences. Not all studies can be completed, and hence focused attention and prioritization of what to study are very important. Theory points to what we might or indeed ought to look at.
4. Translation and extension of knowledge to the world, i.e., beyond the laboratory, is invariably a goal of areas such as clinical, counseling, educational, organizational, and other areas of psychology where theory, research, application, and practice are all important. The best way to advance application is through understanding how something operates, i.e., what are the critical mechanisms? Understanding how and why something works can be used to optimize the effects of a particular influence. For example, there is now a keen interest in seeing if various forms of treatment, well

studied in laboratory, can be effective in clinical practice. Unfortunately, there is very little knowledge of why and how treatment works, so we really do not know precisely what to extend to clinical practice, what ingredients of therapy are necessary, sufficient, and facilitative. Without understanding, interventions are going to be difficult to extend in a way that will be very effective or at least optimally effective. We will not be sure what to emphasize and why and what is essential to include and what can be let go or omitted.

4.7.1: Some Additional Reasons Why Theory Is Needed

In vastly different context—well maybe not that different from treatment—security blankets, small stuffed animals, pets, and parents can comfort very young children in stressful situations. For example, in one experiment, with 3-year-olds undergoing medical procedures, security blankets and moms were equally effective (compared to no supportive agent) in reducing stress and providing blankets and moms did not surpass the benefits of the separate support source (Ybarra, Passman, & Eisenberg, 2000). It would be very informative to understand a range of processes (e.g., biological and psychological) that are involved in southing a child. It may be that people in general can be comforted in several ways and understanding the different ways and commonalties in how they operate would require theory and research. The knowledge once gained might well have broad implications for allaying fear, addressing loneliness, and teaching coping, in relation to children but adults as well. We might, for example, have many different ways of comforting individuals. It would be useful to know if some are more effective than others and whether there is an optimal way of matching source of comfort (e.g., a decadent chocolate dessert, meditation, warm showers) based on knowledge of moderators—what source for what type of person or setting?

Returning to moms and security blankets, most of us probably believe that there are circumstances in which moms are "better" at allaying children's fears and stress. It would be useful to theorize why and then under what circumstances moms are better than blankets. There is more here than just comparing blankets and moms but understanding similarities and differences among comforting influences. Without more research, one would not want to make a "blanket" statement that the influences are the same.

Overall, the goal of science is to understand and this entails connecting empirical relations with statements of mechanisms and process. We do not only want to know that the universe is expanding but to understand how and why. There may be implications for understanding our origins better but also for drawing on novel resources (e.g., for energy, light). Similarly, we do not only want to know that most crime among adolescents is committed while youths are under the influence of an illicit substance (e.g., alcohol, drugs), but why. It may be simply that inhibitions are reduced and restraints that thwart lawbreaking are reduced, but it may be other influences as well such as selection (those who abuse illicit substances are more likely to commit crime whether they use the substances or not, or peer relations in which substance use occurs foster crime, and so on). The value of understanding is critically important, in this case, to intervene to reduce or possibly prevent the problem.

4.7.2: Generating Versus Testing Hypotheses

In beginning a single study or a research career, investigators often are encouraged to start with a theoretical statement or model of the variables of interest and then to test the model empirically. Testing hypotheses based on a conceptual view is sometimes seen as the better and indeed the only way to develop and conduct research. However, this emphasis raises an immediate dilemma.

Where does one get a conceptual view to begin with?

Clearly, there is no substitute for brilliance and keen intuition for generating explanations about why things are the way they are. Also, there is also no substitute for close-up observations and contact with the phenomenon of interest. Meeting, working with, and participating in the situations or contexts one wishes to understand generate reams of ideas about what needs to be studied, what is really interesting, and what processes are involved. Obviously, if one is interested in domestic violence or suicidal ideation and attempt, it is important to work with individuals directly who experience these circumstances and conditions. Observing can be very helpful if for no other reasons than dispelling stereotypes that may have led your research astray or more nuanced identifying stereotypes that have a strand of truth that ought to be clarified. I mentioned the importance of the case study previously as a way to generate research ideas. Contact with the phenomenon of interest is the same point whether one or two cases or exposure to a setting.

Qualitative research is a methodology not taught very much in undergraduate or graduate programs in psychology in the United States, but that is quite relevant as a source of ideas and theory. A characteristic of qualitative research is to conduct in-depth or intensive interviews of individuals and groups who experience a particular situation or show a special characteristic. From such interviews, one can develop in systematic ideas about what are key dimensions of a problem and what needs to be studied. In qualitative research, the term *grounded theory* is used to

denote that hypotheses emerge from intensive observations of the phenomenon, i.e., theory comes from and is grounded in observation. I mention the issue here because it is easy to say here that ideas will flow once one works with the phenomenon of interest. It is likely that this is too nebulous to be of much help. However, there are systematic ways in qualitative research to speed this process by meeting with individuals and groups with special experiences of interest to the investigator and to move from description to explanation.

4.7.3: Further Considerations Regarding Generating Versus Testing Hypotheses

Within psychological research, often there is reluctance in interviewing or chatting with subjects in formulating research ideas. This is understandable because psychological research utilizes many different animals (e.g., nonhuman primates, pigeons, mice, rats, *Caenorhabditis elegans,* reptiles, fish, dolphins, bats, foxes, voles, drosophila, spiders, honeybees, leeches, crayfish, snails, and cockroaches).[9] For most of these, having a focus group or chatting about how they experience phenomena of interest may not be informative. Also, for so many topics (e.g., perception, memory) and dependent variables of interest (e.g., different types of neuroimaging), subjects may not be able to report on the topic of investigation (e.g., "Hi, I wanted to chat with you about what parts of the amygdala might light up when I ask you to imagine . . . ."). With the obvious out of the way, it still may be important to communicate with individuals who experience the phenomenon of interest. Humans often cannot report on influences or reasons guiding behavior, but it is often useful and meaningful to listen to what they have to say to direct one's attention to questions or topic of interest. Qualitative research is a very systematic way to do this, but less formal focus groups and interviews can be helpful too.

Within psychology, purely descriptive research that is not guided by a strong conceptual view is often looked at negatively at worst or ambivalently at best. There is some basis for concern about research that might merely study the relation of any two (or more variables) whether or not the case is made that these variables and the relation are important or have implications for anything. For example, one could study length of one's hair and propensity for depression, blood pressure and shoe size, and attitudes toward government and one's toothbrushing habits (this last one was my undergraduate thesis, I might add). The rationale might be that one is merely describing a relation to generate a conceptual view. In the extreme, any line of work can be made to seem odd. Clearly, there needs to be some basis that the phenomenon of study has some interest and that the study, if not based on theory, might well be useful in generating relations that would be informative.

Some examples might make the point.

What happens to individuals when they drop out of therapy very early?

In my own work, there was no strong theory to pursue this question or to make predictions. As might be expected, many people who leave therapy early are doing poorly, i.e., have not changed appreciably in the clinical problems that brought them to treatment. Describing who leaves early and improves and who leaves early and does not might well generate some interesting data about therapeutic change and attrition. This is not a theory-based line of work but could and eventually did lead to some theory about who drops out of treatment and who stays in but profits less (e.g., Kazdin, Holland & Crowley, 1997; Kazdin & Whitley, 2006). The work began with interest in describing a problem (dropping out of treatment) and evaluating different outcomes (who still gets better and who does not) and from that the possible reasons why.

As another example, among individuals who have a heart attack, those who experience depression are more likely to have another heart attack and to die from it (e.g., Glassman, Bigger, & Gaffney, 2009). Descriptive information about those who have a heart attack and depression but do not have a second heart attack or those who have a heart attack, no depression, and who do have a second heart attack could be quite informative. Moreover, such research beginning purely at a descriptive and correlational level can readily prompt hypotheses that go beyond the data and build theory. For example, mechanisms (e.g., biological processes, coping processes) that connect depression and heart attack are likely to emerge from such studies. If we could, for example, look at one group of individuals (those who have a heart attack and are depressed) and follow the outcomes (those who have a second heart attack and those who do not), we have the descriptive beginnings of potential factors (protective factors) involved. Now we try to explain those protective factors (mini-theory) and come up with some tests of our theory. In an ideal world, we would identify some factors we could manipulate to see if we can reduce the risk of a second heart attack. It is all in here—like a research salad with correlates, moderators, mediators, and a little theory and now garnish with some parsimony (parsley) and we have contribution to basic and applied research as well as a research career.

The goal of research is to understand and theory plays a central role in bringing together multiple variables and processes. Although it is important we end up after several studies with an explanation of what is operating and why with a given topic, we need not start with a conceptual view. Stated another way, we demand of most researchers that they begin their write-up or

article with a conceptual statement (a model or theory) followed by predictions. It would be equally useful perhaps to light this demand and make it more flexible so that researchers must either begin or end their write-up in that way. Research that attempts to describe and to generate hypotheses for further research might not begin with a theoretical statement and predictions. However, at the end of the article (Discussion section) the study might well connect what has been found with some theory and make predictions that are followed in a second study.

Good data on a topic are a great basis for developing theory and a key to understanding. Indeed, it is easy to see occasional examples of theory-based research where the information on which the theory was based was so removed from reality or where the person derived the theory in his or her office with quite little contact with the world. Even so, in many cases where mathematical models are used to describe and then generate predictions, that actually works well. The variables, world, and predictions are represented symbolically and derivations are made from them to make predictions about the world. This is not that common in clinical psychology but to be encouraged.

The interplay between theory and empirical research and between explanation and description is reciprocal. I have noted this section as generating *versus* testing hypotheses to indicate a tension in the field. However, a good study can do either and often a superb study does both. That is, a hypothesis (theory prediction) may be tested, but the data gathered are used to extend the description of the phenomenon in ways that beg for further theory. Alternatively, a study may begin with a careful description and end with a model or conceptual view that can be tested in subsequent studies.

4.8: What Makes a Research Idea Interesting or Important?

4.8 Analyze the causes that make a research idea interesting or important

The emphasis of the chapter is identifying a research idea as the basis for a study. As I mentioned, this can be daunting. It is worth commenting briefly on the quality of the idea or basis of the study because this is extremely helpful in selecting among the infinite possibilities of empirical relations to test. The guiding question of this section as to what makes a research idea interesting or important is easy. There are two overlapping ways to answer this.

4.8.1: Guiding Questions

The type of question one asks can influence whether the research is interesting or important. Type refers to a higher level of abstraction that the very specific hypotheses and variables one is asking.

For example, a study is likely to be interesting or important by the extent to which it addresses:

- Why or how does the question guiding the study represent something that is puzzling, confusing, or perplexing in some way (e.g., the effects are sometimes found but sometimes not and this study might show why or how)?
- Does the study represent a challenge for the field in any way (e.g., to show the phenomenon, to measure something that has been difficult to measure)?
- Could the research finding alter people's thinking on the topic (e.g., is music beneficial for physical health, does something seemingly good [nurturing] have any negative effects, and does something negative [cigarette smoking] serve any positive outcome)?
- Can the research begin a new line of work (e.g., studying grandparent diet and physical health of their grandchildren—this is not too new, but what about grandparent upbringing and diet on the mental health of their grandchildren)?
- Does the research advance or require a new explanation (theory) (e.g., we believe that cognitions play a critical role in depression but maybe not or maybe not in a subtype of depression)?

These questions are useful to mention because they convey what is likely to make a study interesting (see Arguinis & Vandenberg, 2014). They also convey that the bar is high and perhaps unrealistically high for designing a study and for evaluating a study one is reading. After all, each study cannot be groundbreaking. Even so, it is useful to know approximate targets, which the above questions reflect. Even if one cannot always hit the target, aiming one's bow and shooting the arrow in the right direction is probably wise.

You Do. A second guide to making research interesting integrates some of the above but is more realistic and practical.

What makes a study interesting and important? The answer is "you." But a little more is needed to explain this.

Researchers beginning in an area or trying to persuade someone (e.g., advisors, journal editors) that the study is important often are quick to note that this study is the first time something has been investigated. A rationale that something has never been done before is not a very good case for a study because all sorts of "wild and crazy and worthless ideas" (to quote comments from a member of my dissertation committee) can be "firsts." Firsts are not so important, at least for that reason alone. (My proposed philosophy thesis on why Smokey the Bear,

Atilla the Hun, and Peter the Great shared the same middle name [the] was definitely a "first" but hastily rejected as not sufficient.)

4.8.2: More Information on Generating Guiding Questions

> A research idea is important if it can be shown to answer a question that is important, to fill a gap that is important, to test some theoretical proposition, or to cast something in a new or different light.

For example, one can make the case that soldiers returning from war who seem fine might develop PTSD.

What do we know about that, what is missing information, and why is that missing information important? If these can be answered, the study or research idea may well be important.

An idea that may be viewed as an important contribution to the literature often involves focusing on a problem area or unresolved issue in the specific research area of interest to the investigator. To develop a study on a problem or unresolved aspect of a given literature, detailed knowledge of that literature is extremely helpful. There is simply no substitute for knowing the area thoroughly. Reading incisive reviews and individual studies from the relevant literature is helpful; writing such a review may even be better. Although there is no substitute for expertise to generate a research idea that takes an important "next step," mastery of the literature can be delimiting as well. The literature in a given area reflects a set of agreed-upon assumptions and methods, many of which are accepted on faith. Drawing upon areas outside of the content area to be researched frequently adds new dimensions that might not have been pursued otherwise. Thus, the advantage of novice researchers often is that their thinking is not confined by the standard topics, procedures, and methods that have come to be rather fixed—some for good reason, but others from tradition.

I noted that what makes a study important is "you" because it is important to take the reader of a proposal or publication through the steps to make the case logically and cohesively; that the topic, study, and foci are important; and that the specific issues being studied ask critical questions. This requires knowing the context of one's study, what other research has shown, and what is missing that is critical to advancing knowledge.

If one is beginning a research career or this is one's first project, the focus ought to be on a feasible project that allows one to do all the steps simply as outlined below. Rather than trying to play a concert piece or hit a home run with a head turning, Nobel laureate-type idea, research begins as a shaping process where one does the steps and project get more intricate, nuanced, and so on as one's skill develops. It is important to know some of the criteria in determining whether an idea is important—that will be helpful in designing as well as reading studies. Yet, it is important—arguably more important—to master the skills set of developing and writing up one study.

4.9: From Ideas to a Research Project

4.9 Report the importance of the right idea for a research project

Deciding what to study and generating the idea for research can be the most difficult challenge for individuals starting out. We have discussed many avenues to prompt that idea for what one wants to study as well as key concepts that can guide the focus of a study.

4.10: Overview of Key Steps

4.10 Review the steps and decision points to follow when progressing from research idea to project

Once the idea is in hand, there are of course other steps to move this from something general to a research project. I highlight a few of these to place into context the movement from an idea to a research project.

> The process of designing and conducting research project consists of a series of steps and decision points. One can easily identify the beginning and end steps for a project as, for example, reflected in identifying the research idea as a first step and eventually writing up or reporting on the completed project as a final step.

There is a way in which there are steps to a research study as various tasks unfold over time and in a sequence. Obviously, one must have an idea for the study and that precedes collecting data and then analyzing the data. Yet, there is another way less obvious in which the steps are not in a sequence but are important to work out in some way all at once at the beginning of a study before any subject participates. For example, ethical treatment of the participants and how the data will be analyzed are facets of the study that are considered at the design stage. In other words, identifying the idea but also making it into a researchable project all emerge at the beginning of the study. Developing a study usually requires a proposal that is reviewed by an Institutional Review Board or investigation committee. This is a group charged to evaluate the proposal and that evaluation may include all facets of the study. The reason is that the various facets of a study are interrelated.

For example, if the research question is not very important, or if the design is not suited to the research question, or if the sample size is way too small, then ethical issues emerge. Should the subjects be inconvenienced in any way (even just filling out measures) if the study is poorly planned and developed? Review committees are apt to say, "no." The point here is merely to convey that much of a study needs to be worked out before any subject is run because it is wise to do so from a methodological standpoint and because it usually is required.

4.10.1: Abstract Ideas to Hypothesis and Operations

The general idea now must move to something more concrete and testable. Essentially, one has to identify exactly what is being predicted. These can be stated as hypotheses or expectations of what will happen. This is not minor. The challenge is twofold:

- One must make the idea so that it can be tested. That is the hallmark of a scientific hypothesis.
- Once expressed in a testable form it must be an idea that can be supported or refuted by the results of a study.

What would be a finding that would be consistent with the prediction you are making?

If people who are *x* (e.g., are diagnosed with bipolar disorder), do *y* when confronted with a challenge that would be consistent with my little theory (explanation). In science we do not prove theories quite this way, but we make predictions that could be tested and with outcomes that would be consistent with one's prediction.

As if not more importantly, what result would occur that would challenge my theory?

This is critical. Falsifiability has been often considered a main criterion in scientific research. We cannot unequivocally prove a theory. There may be other explanations. But we can falsify them a bit more easily. For example, it is almost impossible to prove the assertion that "all methodologists are extremely happy." We would have to find and test everyone (including those from the past and those not yet born). It is easier to falsify this—all we need to do is to find one unhappy methodologist.

For our theory, we look for ways to test but also to see if the theory stands up when a possible challenge occurs that might require rejection or modification of the theory. That is, we can find something that perhaps disproves the theory or makes the theory in need of modification. There was skeptical excitement a few years ago in relation to the theory of relativity. The theory holds that nothing is faster than the speed of light; then some scientists found that traveling neutrinos (subatomic particles) in fact were traveling faster. Headlines hit the news; many tests were run and rerun to check. If the data were accurate, this would require a major modification of the theory and actually many of its facets. Alas, Einstein's view survived. A nuance of assessment in tracking speed gave the appearance of travel that was faster than the speed of light. Of course, this is science so the topic is not necessarily closed. For now, the speed of light still holds as the limit. The excellent feature of all of this was that in principle there are tests (many actually) of the theory of relativity that could refute key components.

> It is essential to include in one's view about a study what it would take to provide evidence inconsistent with that view. The reason is that we do not want a squirmy theory that can accommodate seeming exceptions or any finding no matter how it came out.

In everyday life, this kind of thinking is not required. So, for example, one can say that a person is "passive-aggressive" and that usually means they said no to something or did not do the expected and they were "really" expressing aggression. The difficulty is that passive-aggression can explain almost all behavior and is difficult or almost impossible to refute. Or when something happens that is clearly inconsistent with one's view, we can say, "The exception just proves the rule." This is all fine in everyday parlance perhaps, but we cannot have slipperiness with scientific hypotheses. So if the finding comes out opposite from one's theory, we would not want to be able to account for that no matter what happened. That is, for a theory to be a scientific theory it must be able to be refuted or corrected.

4.10.2: Moving to Operations Constructs and Procedures

The move to hypotheses is a step toward being more concrete about how the idea will fit into an investigation. The hypotheses will include constructs (concepts), and we need to move toward making those concrete too. For example, one might ask at a general level such questions as, "do anxious people tend to withdraw from social situations?" or "are college students put to sleep by the lectures of their instructors?" Now we look to key concept (e.g., anxiety, social situations, and even "sleep") that seems so straightforward. For research, a lot more is needed to bring these into the realm of a scientific study.

The concepts included in the abstract notion must be operationalized, i.e., made into operational definitions.

> Operational definitions refer to defining a concept on the basis of the specific procedures and methods ("operations") to be used in the investigation.

For example, an operational definition of anxiety could be physiological reactions to a galvanic skin response measure of skin resistance and an individual's self-report of being upset, nervous, or irritable in several situations.

Greater specificity may be required than noting the measure. For example, a study might require operational

criteria for designating anxious and nonanxious individuals. "Anxious" may be operationalized by referring to persons who attain relatively high scores (e.g., at or above the 75th percentile) on a standardized measure of anxiety. Nonanxious or low anxious persons might be defined as those who attain relatively low scores (e.g., at or below the 25th percentile) on the same scale. Specifying the measure and the cut-off criteria to define anxious and nonanxious groups would clearly satisfy the requirements of an operational definition.

Consider an example of an operational definition as well as some of the issues they raise. Sexual hooking up refers generally to "brief uncommitted sexual encounters among individuals who are not romantic partners or dating each other" (Garcia, Reiber, Massey, & Merriwether, 2012, p. 161). Such encounters are extensively portrayed in the media in the United States (e.g., movies, television, and best-selling songs and books). Approximately 60–80% of college students in North American colleges have had some sort of hook-up experience. That can involve a range of activities (e.g., kissing and touching, vaginal, oral, and anal sex). The percentage drops to the mid-30s when vaginal or oral sex are used to define hooking up.

> Different ways of defining hooking up are used in research such as casual sex, a sexual relation that lasts only one night, sex with someone whom one has just met (excluding previous friends and partners), and sex when the motive is not related to love, procreation, or commitment (see Garcia et al., 2012).

All of the variants are defensible as operational definitions, and usually are spelled out in more detail than I have provided here. Three points deserve emphasis:

1. In science it is essential to provide the definition of one's constructs, how they will be measured, and cut-off scores or criteria for meeting the definition if relevant. These are essential for interpreting and trying to replicate the results.
2. For most constructs we study, there is no definitive, true, or single definition. Depression, high self-esteem, self-control, conscientiousness, disgust, emotion regulation, and add your favorite constructs all have been studied extensively. They vary in their operational definitions, and it would be difficult to argue that one is or is not the correct definition.
3. Because operational definitions can vary across studies, it is easy to understand why results might not be identical. In many areas of work, there are long-standing measures and they are used to operationally define a construct. For example, to study depression in adults, familiar measures (e.g., Beck Depression Inventory, Hamilton Rating Scale for Depression) facilitate comparison and combination across studies. Reliance on such measures brings slightly more homogeneity to operational definitions across different studies. That is, after decades of use, individuals in an area of research have a great idea about how depressed individuals were in a given study using these measures. Although hooking up is an active area of research, there are no standard measures and definitions to the same degree there are in a more heavily researched area of study such as depression.

   > Selecting measures to evaluate constructs is a key task in developing a research project. The measures define the constructs, and it is important to ask oneself, "Does this measure get at the construct in the way I wish? Is there a better way of defining the construct?" The support for a given hypothesis may vary depending on how one defines the construct, so this is not a trivial task.

A full description of the operational definition is needed for key constructs. Related, the procedures the investigator will use need to be fully and explicitly described. The procedures refer to what happens when the subject arrives at the study from start to finish. Who meets the subject, and what does the subject do? When is the consent procedure presented, what exactly is the experiment manipulation or experience, who administered or provided it, how long did it take, how many sessions were there then, and so on? If the study is done online, the equivalent would be specifying what the subject will be doing, exposed to, and in what order. These are specific to the study and some points I mentioned are relevant and others not, and some points I have omitted are relevant. These can be seen from the method section of articles of published research. The main point: what the investigator does in a study ought to be transparent and explicit.

4.10.3: Sample to Be Included

A critical decision is whom to include as the subjects. Much of research is conducted with college students, and I have noted previously that there are reasons to question the generality of results in evaluation of even core psychological processes (learning, perceptions). The heavy reliance on college students is complemented increasingly by recruiting subjects from online sources (e.g., MTurk, Qualtrics) and that tends to be a sample older than college students and more diverse in their education, occupations, and stage of life.

The issue at the proposal stage is to consider the matter explicitly.

Why is this subject pool going to be used?

- One answer is that one believes the principle or concept one is testing is not likely to be influenced by which among the available samples I select.
- The more usual answer is one of convenience, i.e., subjects were selected because they could be obtained easily or within some time frame. That is the one to be careful of.

The goal is always to provide the strongest test of one's hypothesis. That means, *what are the best circumstances in which this hypothesis is likely to be supported?* The answer to that has many facets, including the measures (operational definitions) but also the sample (that will be used). The sample issue may be related to providing a strong test or to external validity. For example, different parenting and child-drearing practices may be surveyed and related to work experience and marital and family relations.

Is a college student sample the "best" or the most credible? Most college students have not yet been parents, have not been in the workforce, and do not have a full-time, live-in partner who also is in the workforce. It may be that one wants a sample not exposed to the conditions of interest, but it may also be that it is hard to make the case for a college student sample for key issues among these variables. It is important to make the case that the sample provides a fine, reasonable, or great test of the hypotheses.

Often the goals of a research project in clinical entail the use of a special population (e.g., to evaluate cognition in patients with dependence on alcohol or drugs, to follow children who have been neglected to see their varied outcomes in young adulthood). Yet even when that is not the case, it is useful to ask oneself:

- Is there a special population that is really a great test of my hypotheses?
- Stated more crassly and strategically, is there a special group that is very likely to show the predicted results?

One wants to begin research with the strongest test of a hypothesis. That can begin with careful thought to who will be selected to participate.

4.10.4: Research Design Options

The research design refers to the arrangement or ways to arrange conditions to evaluate the hypotheses. There are a variety of options for research related to how the idea is evaluated and the conditions in which the study is conducted. The options have implications for diverse threats to validity and hence the investigator's conclusions. The different ways in which the study might be designed will be discussed later. Here is a preview of major categories to have in mind.

> Research in psychology actively draws upon three major types of studies: true experiments, quasi-experiments, and observational designs. Each of these is a family of designs with many options.

True experiments consist of investigations in which the arrangement permits maximum control over the independent variable or manipulation of interest. The investigator is able to assign subjects to different conditions on a random basis, to vary conditions (e.g., experimental and control conditions) as required by the design, and to control possible sources of bias within the experiment that permit the comparison of interest. From the standpoint of demonstrating the impact of a particular variable of interest, true experiments permit the strongest basis for drawing inferences.

> *A true-experiment is a generic term to apply to studies in which subjects can be randomly assigned to conditions and the investigator controls who receives and who does not receive the experimental manipulation or intervention.*

When true-experiments are conducted in the context of an intervention (treatment, prevention, education), they are referred to as *randomized controlled trials* (or RCTs) and sometimes randomized controlled clinical trials (still RCTs). The term is used in many disciplines (e.g., psychology, psychiatry, education, epidemiology, and medicine) and refers to an outcome study in which clients with a particular problem are randomly assigned to various treatment and control conditions. In clinical psychology, the now vast research on evidence-based treatments has relied very heavily on RCTs. It is useful to be familiar with this term because of its widespread use and because this type of study is recognized to be the best and most definitive way of demonstrating that an intervention is effective. RCT often is referred to as the "gold standard" to convey its special status, but as any single method it has its own limitations.

Occasionally an investigator cannot control all features that characterize true experiments. Some facet of the study such as the assignment of subjects to conditions or of conditions to settings cannot be randomized.

> Quasi-experiment refers to those designs in which the conditions of true experiments are approximated.

This could mean that random assignment is not possible because groups are preformed or that random assignment could be used for some groups but not all (e.g., a control group was added for comparison purposes and that group was preformed).

For example, an investigator may be asked to evaluate a school-based intervention program designed to prevent drug abuse or teen pregnancy. The investigator wishes to use a nonintervention control group because the passage of time and influences that are occurring during that time (e.g., history, maturation, testing, and other internal validity threats) can lead to change. However, for practical reasons a control condition is not permitted within the school that wishes the program. The investigator seeks other schools that will serve as nonintervention control groups and uses students in these control schools for comparison purposes. These other schools might be similar (e.g., in population, size, geography). We have lost the central feature of true-experiment, random assignment to groups, and a host of factors (e.g., motivation for change among administrators) that may differ greatly across conditions. Already the design is less ideal than one would like. Yet,

there are many design options and methods of drawing valid inferences. Quasi-experiments can provide very strong bases for influences and ought not to be ruled out.

True and quasi-experiments refer primarily to studies where an independent variable is manipulated by the investigator, as illustrated by providing treatment or an experimental condition to some persons but not to others. A great deal of clinical research focuses on variables that "nature" has manipulated in some way, as mentioned in the discussion of subject variables earlier in the chapter.

4.10.5: Additional Information Regarding Research Design Options

> Observational designs refer to a variety of arrangements in which the variable of interest is studied by selecting subjects (cases) who vary in the characteristic or experience of interest.

For example, the investigator might wish to study differences between cigarette smokers and nonsmokers in relation to some personality traits or background characteristics, between marital partners who are in the same occupation versus those who are not, between males and females, and between persons who were former prisoners of war and those who were not. Of course, the investigator does not manipulate the independent variable, but does identify groups that vary in the characteristic of interest.

A comparison group or groups are also identified to control for factors that may interfere with drawing conclusions. Observational studies can provide critical insights about the nature of a problem, characteristic, or experience, as I shall discuss at greater length later.

Each of the above type of design is based on studying groups of subjects. Each group usually receives only one of the conditions. Occasionally, the general class of designs is referred to as *between-group research* because separate groups of subjects are formed and ultimately compared. A between-group design includes at least as many groups as there are experimental and control conditions. (Yes, I know what you are thinking and it is true—if there were more than two total groups in the study, it should be called among-group rather between-group research. I am just the messenger and between groups is the term that is used.)

In addition to group designs, clinical research also entails *single-case experimental designs*. These are true-experimental designs but can focus on a given individual, a few individuals, or one or more groups over time. The underlying approach toward research for group and single-case designs is the same, namely, to implement conditions that permit valid inferences about the independent variable. However, in single-case research, this is accomplished somewhat differently. Typically, one or a few subjects are studied. The dependent measures of interest are administered repeatedly over time (e.g., days or weeks). The manner in which the independent variable is implemented is examined in relation to the data pattern for the subject or group of subjects over time. Single-case designs can play a special role in clinical work where the focus is, of course, on the individual client. Single-case designs can be used to experimentally evaluate the impact of a given intervention or multiple interventions. As with group designs, there are many different single-case designs, with their own requirements, advantages, and obstacles.

4.10.6: Multiple Other Decision Points

There are other tasks and decision points in developing the research. A few critical points are noted here as a preview of more extended discussions later.

> **Data Evaluation:** How will the results be analyzed is something to address at the beginning of a study when a proposal is being prepared.

Among the issues, how many subjects is related to statistical evaluation, as discussed as part of data-evaluation validity. Also, it will be important to specify the statistical analyses to evaluate the hypotheses. If this is the hypothesis, what test or comparison would provide the needed support. This is critically important to consider at the design and proposal stage.

Data analyses cannot be completely planned in advance of a study. Many things can happen (e.g., loss of subjects, intriguing, and unexpected ancillary findings) that will lead to different and additional analyses from those that are planned. This is a given for much research. Even so major analyses that are likely to be done to test the hypotheses still should be specified at the outset and proposal stage.

> **Time Frame for Research:** Research often varies in the time frame for investigation. The bulk of research is conducted in a concurrent time frame in which the experimental manipulation (independent variables of interest) and the measures to evaluate the impact of the manipulation (dependent variables) variables are administered and completed within a relatively brief period—in fact usually one session.

An example would be a laboratory experiment in which subjects are exposed to an independent variable and complete the measures that same hour or day. In contrast, the investigation may be conducted over an extended period of, say, several years. A frequent distinction is made between cross-sectional and longitudinal studies. *Cross-sectional studies usually make comparisons between groups at a given point in time. Longitudinal studies make comparisons over an extended period, often involving several years.* Each type of study has special strengths and weaknesses we shall discuss later.

Ethical Protections of the Subjects: At the outset of a study, it is important to identify what special protections are needed and will be provided for the participants.

Will there be any deception, are their potential side effects of the procedures or measures, how will potential side effects be monitored, and what will be done if those emerge? Seemingly innocuous procedures require close attention, and regulations (federal laws and regulations in the United States) about protecting subjects are important at the planning stage of the study.

For example, questionnaires may be administered to college students as part of a study that is innocently correlating variables. The measures may include something on depression, including the standard and familiar measures such as one of those I mentioned previously. The ethical issue: what will be done if a subject endorses items that suggest he or she is contemplating or at risk for suicide? How will the study be monitored so that this can even be detected right away, what will be done if someone appears at risk, and what exactly is the criterion for deciding "at risk" in the study? I raise this to note that protection of participants goes well beyond the obvious concerns about subjecting them to onerous or stress-producing experimental manipulations. What information can emerge that indicates there is a problem, danger, or immediate concern of a subject? In addition, are there any features of the procedures that will be bothersome, annoying, or painful or place participants at risk for untoward mental or physical health consequences? Finally, how will the privacy of participants be protected if sensitive information is collected? How all of these situations and circumstances will be handled, to the extent they can be anticipated, needs to be specified at the proposal stage.

The conditions of the experiment are described to participants; and in most studies where the identity of participants is known, informed consent procedures are administered. Subjects must sign consent forms conveying that they are informed of the conditions and their rights (e.g., to withdraw without penalty, to not answer specific questions). Consent often is not required in subjects where the identity of participants is not known or the information for a study is from archival records (criminal, medical, educational). There are both ethical and legal issues that guide consent.

4.11: General Comments

4.11 Summarize the steps that lead to a successful research project design

I have mentioned some of the major steps and decision points. I mentioned here along with generating the research idea because seemingly distant concerns (e.g., how will I analyze the data, how will I ensure that ethical considerations and matters like informed consent are handled) are present at the very beginning of the study when the plan is made on what to do to test this idea one has generated. Indeed, it is usually the case (e.g., universities) that research cannot proceed until a review of the proposal ensures that critical issues such as those I have mentioned are resolved.

Of course at this early point in designing the study, this is the perfect time to pull out your list of threats to validity (which hopefully you have had laminated with a small credit card size version for your wallet/purse) or set as background on your smartphone, tablet, and laptop.

- What threats are likely to emerge from the study I am planning and what can I do about it now?
- Also, what do I want to say at the end of this study if the predictions I make are accurate?
- Will there be plausible rival hypotheses that compete with my interpretation?

When one answers this at the outset of a study, the study is likely to be greatly improved. As we move to the discussion of specific designs, ways of addressing threats will emerge again.

Summary and Conclusions: Ideas that Begin the Research Process

The research idea that serves as a basis for experimentation may be derived from any of several sources, including curiosity about a particular phenomenon, case studies, interest in special populations (e.g., those who meet criteria for a particular disorder, those with a special history or experience), extrapolation of findings from other types of research (e.g., processes studied in animal research), development of measurements, and many others. Also research may seek to illuminate variables or characteristic are related as reflected in such concepts as correlates, risk factors, protective factors, and causes of a particular outcome. Moderators, mediators, and mechanisms were discussed because they too frequently serve as the impetus for an investigation.

Translational research was discussed as a type of study that moves from basic to applied questions and occasionally

moves back again. The concepts of "bench to bedside" and "bedside to community" refer to types of research that move basic findings to clinical application and then to large-scale application. The process can go in the opposite direction. We learn that an intervention leads to change and go back to the laboratory including nonhuman animal studies to identify processes that might be involved. Translational research emphasizes the movement from laboratory research (bench) to clinical application (bedside) to larger-scale applications (community, social policy). The continuum notes several places that may promote ideas for research.

Whatever the focus, it is to draw on theory to guide a research study. Theory refers to a conceptual view about the relationship, i.e., how the variables of interest relate to each other and to the extent possible, why, and under what conditions. Not all research needs to be driven by theory. Research that carefully describes phenomena too can contribute greatly. The distinction was made between testing hypotheses (e.g., usually theory driven) and generating hypotheses (e.g., describing phenomena so as to generate theory). The goal of research is to understand the phenomenon of interest and theory can help enormously. Yet, one can begin with a theory or end with a theory or more likely both. Both means we have an idea, and it gets modified and enriched based on the data we have obtained. As a research strategy, beginning descriptive work to generate hypotheses and theory or explanatory work to test hypotheses are both legitimate. Both focus on understanding the phenomenon of interest.

Identifying the research idea begins a process with many steps. Several steps were outlined including moving the abstract idea to specific hypotheses and operations, providing operational definitions of measures and procedures, identifying the sample that is suitable or optimal to test the hypotheses, selecting among the many research design options, outlining the data analyses that will be used, and important addressing ethical issues to protect subjects and ensure their rights. Each of these and other such steps is relevant at the very outset before the study is actually begun. Proposals usually are required to identify how these steps will be performed to provide approval to proceed with the study. We will address each of these steps in detail.

The distinction was made of broad types of research designs. True experiments, quasi-experiments, and observational studies were highlighted to define major types of research. Each has many options. Although one type and within that even subtypes, such as randomized controlled trials, are often regarded as preferred, superior, or ideal, questions can and ought to be answered in different and complementary ways. The challenge is to use the best design available to test the hypotheses.

Critical Thinking Questions

1. Moderators and mediators are important topics in research. Give a clear definition of each and then an example (hypothetical or real).
2. Translational research: what is that and what is meant by bench, bedside, and community?
3. What makes an idea for research interesting or important? Name two factors.

Chapter 4 Quiz: Ideas that Begin the Research Process

Chapter 5

Experimental Research Using Group Designs

Learning Objectives

5.1 Review how random selection improves the external validity of experimental results

5.2 Examine the importance of selecting the right sample

5.3 Analyze the importance of selecting the right sample and the right group in research

5.4 Identify the RAX notation used in illustrating the sequence of events in a research design

5.5 Describe the pretest–posttest control group design

5.6 Contrast the posttest-only control group design with the pretest–posttest control group design

5.7 Analyze the pros and cons of the Solomon four-group design

5.8 Express the relevance of the factorial designs when there are multiple variables

5.9 Recognize the areas where the researcher has no control over the subjects as quasi-experimental designs

5.10 Examine the nonequivalent control group designs

5.11 Illustrate how a quasi-experimental design was used to study the impact of secondhand cigarette smoke

5.12 Recognize crossover design as a form of multiple-treatment design

5.13 Identify some of the deliberations that need to be taken into account while choosing a multiple-treatment design

By far the most common research designs within psychology compare groups of subjects who are exposed to different conditions that are controlled by the investigator. The general strategy can entail a variety of different arrangements depending on the groups included in the design, how assessment is planned, and when and to whom the experimental condition is presented. This chapter considers fundamentals of group designs and various options when the investigator manipulates or systematically varies conditions and controls the assignment of subjects to different conditions. We begin with discussing individuals who are to participate in the study, their assignment to groups, and specific arrangements that constitute the experimental manipulation.

5.1: Subject Selection

5.1 Review how random selection improves the external validity of experimental results

A fundamental issue in group designs is the selection of participants for research, i.e., who will serve as subjects in the study? This topic is under discussed in psychological methods because the task of subject selection seems obvious. If one wants to do nonhuman animal research, then the sample (e.g., rats, mice) usually is dictated by the subject matter and whether the animal is a good model for what is being studied; if one wants to conduct a laboratory study with humans (e.g., in social or clinical psychology),

then college students enrolled in introductory psychology or samples available through the Web (e.g., MTurk) may be fine. Yet, there are many issues about subject selection that have great implications for methodological concerns, beyond the obvious matter of external validity or generalizing to a population. Let us consider several issues related to selecting subjects for inclusion in experiments.

5.1.1: Random Selection

Randomness is discussed frequently in scientific research. When investigators discuss randomization in experimentation, they usually are concerned with one of two concepts, namely, random selection of subjects from a population and random assignment of subjects to experimental conditions. In group designs within psychology, random assignment and related procedures to form groups are the central topics and are taken up later in this chapter. Random selection is an independent issue that is not necessarily related to the particular design but warrants mention.

> Random selection refers to drawing from the total population of interest in such a way that each member of the population has an equal probability of being drawn.

If that is accomplished and the sample is relatively large, one can assume there is no special bias in who was selected. Random selection enhances to the generality (external validity) of experimental results.

> If we wish to generalize results from a sample of subjects in the experiment to a population of potential subjects, usually it is essential to select a representative sample of the population.

We would not want to restrict selection to patients in a particular hospital or clinic or in a particular city, state, or country or of one ethnicity but would want to sample from all available persons. If subjects can be drawn from the entire population, it is more likely that the sample will represent the population of individuals who are depressed. Generality of experimental results (external validity) depends upon the representativeness of the participants in the experiment to those individuals who were not included, i.e., the rest of the population.

There is an obvious restriction in principle as well as practice to random selection. Subjects in an experiment cannot be selected from a population unless that population is very narrowly defined. For example, for a population defined as "all introductory psychology students currently enrolled in a given term at this university," a random sample might be obtainable. However, a random sample of "introductory psychology students in general" could not be readily obtained. To sample this latter population would require being able to select from all individuals who have had introductory psychology already, including those no longer living, all those currently enrolled, and all who are yet to enroll (including unborn individuals) across all geographical settings.

> Sampling from all subjects in the population including those who are deceased or yet to be born, of course, is not possible.

What that means is that a finding obtained by a sample at one point in time may not generalize to the population from different points in time. Staying with just the present and all living subjects, generality of the experimental results to a population depends upon having randomly sampled from a population. Without that, conclusions would seem to be restricted to the narrowly confined groups of subjects.

5.1.2: More Information on Random Selection

Random selection from a population is often central to research. For example, epidemiological research identifies the distribution of various conditions (e.g., diseases, mental disorders) within a population.[1] In such studies, special sampling procedures are used to ensure that population of interest is represented. An example familiar within the mental health professions is the research on the prevalence of mental disorders. In the United States, approximately 25% of the adult population meets criteria for at least one psychiatric disorder at a given point in time (Kessler & Wang, 2008).

> To make statements about the population, careful sampling is required of different segments or subgroups of a population to reflect subject and demographic variables of interest, such as geography, socioeconomic level, ethnicity, religion, and other variables.

Within such groups, persons are selected randomly so that the final sample reflects this distribution. In survey and opinion poll research as well, sampling from the population in this way also is critically important to ensure generality of the findings to the larger population, within some margin of error. Also, it may be of interest to divide the data by the various subgroups in the sample and report data separately. For example, in surveys, the views or responses are often separated to compare women versus men, younger versus older, and those of various ethnic or cultural subgroups.

In psychological research, whether in clinical, developmental, or other specialty areas, random sampling from a population is rarely invoked. A representative sample is not considered as essential nor is its absence viewed as an issue. There are many exceptions. For example, in one program of research on health, the focus is on a single county in the State of California. The sample of approximately 7,000 was selected from the population of individuals in

that county, and that sample was followed for a period spanning approximately three decades.

> Among the interesting findings is that subjective well-being relates to physical health. Subjective well-being includes global satisfaction with one's life, satisfaction with specific domains of one's life, positive affect, and low negative affect.

Higher levels of subjective well-being were associated with lower death from both natural and unnatural causes over a period spanning 28 years (Xu & Roberts, 2010). The effect of subjective well-being was mediated by having social networks. We know from other studies too that happiness measured longitudinally is associated with living longer (Frey, 2011). In any case, this is an example that conveys two points about sampling. First, the study focused on a population and sampled randomly to represent that. Second, population can refer to a well-defined group and does not invariably mean everyone (e.g., in a country, the world). In this case, the focus was on "everyone" in a particular county in one state and reflecting that group in a representative sample.

In any case, populations occasionally are studied in psychological research by random selection of subjects from a well-defined larger group. Yet, this is the exception as I have noted and reading research articles within psychology in virtually all of the journals will reveal samples that were not selected specifically to represent a population. This is not necessarily a problem or restriction of psychological research, but it does call for keeping the concepts and practices of random selection, which is not used very much, distinct from random assignment, which is routinely used.

5.2: Who Will Serve as Subjects and Why?

5.2 Examine the importance of selecting the right sample

If the sample we are to use is not random and cannot be said to represent a population, then why are we using the particular sample we have selected for study? If we are not specifically attempting to represent everyone (in a well-defined population), much further thought needs to be given to who *is* selected and why. There are a few critical issues that ought to be considered explicitly when beginning a study.

5.2.1: Diversity of the Sample

Diversity of the sample is one such issue and has been of enduring concern. For decades, much of the research in the United States (e.g., psychological, biological) was conducted primarily with European American males (see Graham, 1992; Guthrie, 2003). Women and groups of other cultures represent multiple and significant segments of the population, and there was no compelling scientific rationale for their exclusion. Indeed, insofar as findings from research have immediate or long-term implications for physical and mental health and healthcare decisions (e.g., policy, legislation), the absence of research on the various cultural and ethnic groups might even be considered as discriminatory. That is, benefits intended from relying research findings might unwittingly help males more than females.

Spanning over a decade, recommendations have been clear to consider cultural and ethnicity diversity in conceptualizing and conducting research (American Psychological Association, 2002). This includes sensitivity to cultural and ethnic issues in conceptualizing and conducting research insofar as it should not be assumed that these variables make little or no difference or are nuisance variables to be handled by merely assigning all comers to various conditions. In relation to the present discussion, diversity of the sample especially in relation to ethnic, cultural, and sex ought to be addressed in designing a study unless there is a compelling reason to focus on a much narrower group.

In addition to the limited sampling of women and ethnic groups, the extreme reliance on college students further restricts who is included in research. Many findings from college student samples (called WEIRD, as an acronym for individuals who are Western, Educated, Industrialized, Rich, and from Democratic Cultures) do not generalize to others (non-WEIRD people) (Henrich, Heine, & Norenzayan, 2010a, b).

> Fundamental psychological processes (e.g., perception, memory, perspective taking) vary as a function of ethnicity and culture, and these processes are not represented in WEIRD samples.

As researchers, in principle we usually do not want to restrict our conclusions to the one local group we are studying and also to a highly homogeneous group. The default emphasis has shifted from selecting homogeneous subjects for research to including a diverse sample to reflect better multiple dimensions of the population. Even when college students are relied on, diversity has been facilitated by changes in college recruitment and admissions. No student samples are more diverse than they have been in the past.

It is not necessary to select subjects randomly from a population but rather to avoid systemically excluding subjects who in fact are in the population and the diverse characteristics they reflect.

> Sex, sexual identity, ethnicity, culture, and socioeconomic level (usually defined by occupational, education, and income) are merely some of the variables that support the importance of diversity of the sample.

Each of these "variables" could and does moderate all sorts of relations of interest. For example, in clinical psychology such topics as the family, stress and coping, social support, child rearing practices, bereavement, participation, or seeking treatment enter into many areas of research. There are ethnic and cultural group differences in relation to each of these areas, and one ought to ensure that these differences are addressed, revealed, and eventually understood in research.

Similarly, socioeconomic status is a variable that has pervasive influences on all sorts of mental and physical health outcomes (e.g., Haas, Krueger, & Rohlfsen, 2012; Sen, 2012). It is very likely that socioeconomic status will moderate many relations and findings in psychological research. Indeed, on a priori, rather than empirical, grounds one can conclude that socioeconomic differences will play a major role in some areas of research. The base rates of some characteristics psychologists study (e.g., high levels of stress, illness, early mortality) differ as a function of socioeconomic status, and this can influence the correlations of these variables with other influences that are studied. So if socioeconomic status is not of interest in a given study, it may need to be evaluated and controlled to clarify the relations of other variables that are of interest.

5.2.2: Dilemmas Related to Subject Selection

There are dilemmas related to subject selection to be aware of in conducting and interpreting research.

> **The first dilemma** relates to diversity of the sample included in a study. We know well that sampling one ethnic or cultural group may limit the generality of findings even though it does not necessarily do so. Some things are generalizable across subject characteristics, but we are not sure about which ones are or are not if they have not been studied. This argues for including a diverse sample in one's research unless there is a compelling reason not to. Indeed, a particular hypothesis or characteristic may require a very limited sample (e.g., Latina women, adolescents from rural America). However, the default position is to include a diverse sample rather than not to.

The challenge is that diversity is, well, rather diverse! For example, the U.S. Census recognizes five racial groupings (leaving aside combinations):

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or other Pacific Islander

(Race is recognized to reflect social, political, and cultural definition of groupings rather than distinct scientific or biological categories.) Moreover, these racial categories are further combined with two ethnic categories (Hispanic or non-Hispanic) for all the combinations and permutations. As it is, the groupings are hardly satisfactory and arguably not very meaningful. Among the reasons is that broad labels (e.g., Hispanic American) can include multiple groups readily distinguished culturally and genetically. Leaving the United States, we know worldwide there are hundreds of ethnic and cultural groups (www.infoplease.com/ipa/A0855617.html). In principle or practice, we can never evaluate a particular phenomenon or finding in relation to all different ethnic and cultural groups (see Kazdin, 2008a). On the one hand it is important to include diversity in the sample, but on the other hand to recognize that any sample is inherently limited in representing all groups.

> **The second dilemma** expands the issue of ethnicity and culture. There are many moderators beyond ethnicity and culture that can have pervasive influences in the psychological processes or relations they influence. Sex, age, and socioeconomic status, for example, are likely to moderate all sorts of findings. The dilemma is recognizing that these and others moderators may be just as pervasive in their impact as ethnicity and culture but cannot be included in any comprehensive way in a given study.

Gender identity too may be a moderator on equal status with the ones I have mentioned, depending on the focus of the study and hypotheses.

How do you think we should proceed?

In selecting subjects, it is important to have a rationale as to why this particular sample provides a good test of the hypotheses and also to include diverse subjects. The role of theory in research was commented on earlier in generating research ideas. Theory is relevant to subject selection too because it may suggest characteristics of a sample that provide a strong or ideal test of the hypotheses. Typically generality of a finding (external validity) is not the first goal of a study. That argues for providing the strongest or best test (for internal validity). If a hypothesis is likely to be more evident in one situation rather than another and with one sample rather than another, that is quite fine to restrict the study to those situations and samples. However, the task of the investigator is to make explicit why a particular sample was selected.

A final comment on diversity that connects directly to methodology and other topics we have discussed. Diversity in everyday life has its own meanings.

In the language of methodology, diversity has other terms including variation and variability.

Recall that demonstrating an effect can be greatly facilitated by reducing variation and variability. The more variable the sample, for example, the lower the effect size for a given difference between means. Data-evaluation validity

argues for a less variable sample. This chapter argues for a more variable (diverse) sample. How to reconcile? Select the sample that you believe provides the strongest test of the hypothesis. If special sample characteristics are not essential, then represent the people of our culture as best as possible. It is likely that some characteristics (e.g., age, presence of a psychiatric disorder) will be relevant and of course select and exclude on the bases of those.

5.2.3: Samples of Convenience

It is often the case that a sample is selected because it is around or available. A sample of convenience is a set of subjects that are studied because they are present in a convenient situation (e.g., waiting room, hospital ward) or is available for a quite different purpose (e.g., participation in another experiment that requires a special population). An investigator may use an available sample to test a particular idea or to evaluate a measure he or she has just developed. However, in a sample of convenience, often there is no clear rationale as why this sample is important, useful, or relevant to the study.

> College students who serve as subjects constitute the main instance of this. Few researchers (but many parents and university administrators) are really interested in specifically how college students behave, but they are selected because they are captive and in many cases are required or have to complete experiments as part of their participation in a course. Samples that are used merely because they are available are referred to as samples of convenience.

Perhaps because college students are used so frequently, the term "sample of convenience" usually is not applied to students. Similarly, individuals available online who serve as research participants are another sample of convenience. The term "sample of convenience" often is used pejoratively with the implication that you should not have used them or we were simply lazy. Yet, the issue is whether a sample is appropriate and not whether one went to horrible pain and hoops to get the sample.

There is another concern about samples of convenience that deserves further consideration and justification. In this situation, a project is conducted to evaluate a special population (e.g., parents of children who visit a clinic for their diabetes, a sample of psychiatric patients). As that study is begun, the original investigators or other investigators (e.g., students, postdoctoral researchers) realize that the data set can be used to test other hypotheses, even though the original sample may not be the sample that would have been used if these other, new purposes were the central part of the study. So now the new study may be proposed that studies emotion regulation strategies or attachment style and adherence to medication for their disorder (e.g., diabetes). Clever hypotheses are devised and are to be tested, and the data set is convenient so these will be the subjects used. Maybe some new measures (e.g., on emotion regulation, attachment, adherence) will be inserted and the researcher can relate these to the large database already available. Now the question is whether this is a good, reasonable, or suitable test? Is there something about this very special population that actually could influence the variables under investigation (e.g., moderators, confounds)? Is adherence to diabetic treatment (medication, monitoring, injections) like adherence to other treatments? And is emotion regulation or attachment different from what it would otherwise be in a population that perhaps had to control parts of their lives very carefully and in the early years relied heavily on others (parents, medical staff)?

When a sample of convenience emerges in the fashion I have noted, the onus is on the investigator to evaluate or at least discuss whether unique features of the sample may contribute to the results. In this context, the use of a highly specialized population that is selected merely because it is convenient raises concern. It is not clear that the sample is well (or for that matter poorly) suited to the question. The specialized population and the factors that make them particularly convenient may have implications for generalizing the results.

The entire issue of sample of convenience raises a broader question that is pertinent to all research. Some rationale ought to be provided why the sample was selected (e.g., college students, people of one ethnicity, a particular age) for any research project. More thought about the sample could enrich the hypotheses and yield as well. The thought might prompt more hypotheses to test or generate predictions about characteristics (moderators) that influence the findings. On the other hand, the investigator may feel that the population is not relevant. That would be a rather strong claim and would be worth making and explaining. The default positions are to include a diverse sample and to explain why the sample is suited to the question that is studied.

5.2.4: Additional Sample Considerations

In some research, selecting a very restricted sample is fine for several reasons. The goal of the study may dictate a restricted sample. For example, studies of postpartum depression and breast cancer focus on women who experience the problem. The main interest is in women, although each of these disorders is also evident in men. Including men might not be feasible (lower prevalence rates for these disorders) and introduce variability (sex differences) that is not of interest to the investigator. Also, it is quite likely that different processes are involved for males and females in the onset and course of the disorders.

In other research, the investigator may view the sample as not particularly relevant or critical. The demonstration may focus on a phenomenon that is of theoretical significance. Ivan Pavlov's (1849–1936) research on classical conditioning with dogs is of this ilk. It is fortunate for us that Pavlov did not start worrying immediately if dogs of different sizes, temperament, color, age, and weight, not to mention dog socioeconomic status, would also show the effects conditioning. Fortunately as well there was no naïve and annoying methodologist who was peppering Pavlov with questions about external validity. Years later we learned of the amazing generality of the phenomenon of classical conditioning across a wide range of species and circumstances, but this was not the initial import of the finding.

Even when generality is of interest and important for the initial finding, a broad and diverse sample is not always available for study. In most settings, not all the cultural and ethnic groups from which one might like to sample are available. For example, my own research is conducted at a clinic for the treatment of children and families.[2] The clinic draws from local communities and hence includes European American, African American, Hispanic American, Asian American, and combinations. The first two groups comprise over 90% of the sample, and only these groups can be examined in the data analyses. The small numbers have never permitted data analyses of other groups because of inadequate statistical power. Often with single-site studies, there are practical constraints. However, more and more research is conducted at multiple-sites simultaneously, and that allows a broader and more diverse range of subjects to include in the study.

5.3: Subject Assignment and Group Formation

5.3 Analyze the importance of selecting the right sample and the right group in research

Selection of the sample, i.e., who will serve as subjects, is of course quite different from how subjects, once selected, are allocated to various groups or conditions in the study. A fundamental issue of research is ensuring that subjects in different groups or conditions are not different before the experimental manipulation or intervention is provided. Recall that we previously discussed selection (group differences) as being a fundamental bias or threat to internal validity. Selection in this sense does not refer to who the subjects are but rather whether groups may differ because subjects selected to serve in one group differ from those selected to serve in another group. A goal of research is to equalize groups except for the one variable (or multiple variables) that the investigator wishes to study or evaluate.

5.3.1: Random Assignment

Once the group of subjects has been selected for the study, it is critical to assign them to groups in an unbiased fashion.

> Random assignment consists of allocating subjects to groups in such a way that the probability of each subject appearing in any of the groups is equal. This usually is accomplished by determining the group to which each subject is from generating a list of random numbers or looking at a table where such numbers already are listed.

Typically, the random numbers are generated by readily available Web sites but the preexisting tables often are in an appendix of statistics text books.[3]

Let us work out how to do this with a brief example. Consider we are going to conduct an experiment with three groups and we will assign subjects to each group. We will label the groups arbitrarily as 1, 2, and 3. Now we need random numbers that order 1, 2, and 3 several times, with each number referring to one of the groups in the study. We can do this by going to a search engine on the Web and typing in "random numbers generator" and access one of the many programs that allow us to specify the number of groups (in our case 3) and the number of subjects (let us say N = 90). The generator will now give us 90 numbers, where 1, 2, and 3 are in a random order.

Alternatively, we consult a table of random numbers and now enter a column or row and look at all the numbers in order. We draw just 1, 2, and 3, and as we go down the columns or across the rows do this to get enough numbers for our 90 subjects. From either the Web or table our final list would include 90 numbers listed in the random order (e.g., 1, 1, 3, 2, 3, 3, etc.). (Numbers other than 1, 2, or 3 in the printed table in various statistics text books, of course, are ignored.) As the subjects arrive to the experiment, they are assigned to the groups in order according to the number that was drawn. So the first two subjects in our study would be assigned to group 1, the third to group 3, and so on in order. With such assignment, subjects are effectively assigned to groups randomly, according to the predetermined schedule.

Drawing random numbers to determine group assignment does not guarantee that an equal number of subjects would be assigned to each group. In the above example, the number 3 may have been drawn from the table more times than the numbers 1 and 2, and thus more subjects would be assigned to this group than the other groups.

> For power of statistical tests (data-evaluation validity) and convenience in conducting several statistical analyses, it is better to have equal rather than unequal group

sizes. This can be accomplished without violating random assignment by grouping subjects into sets or blocks.

Each set consists of the number of subjects that equals the number of groups in the experiment. If there are three groups, the first three subjects who appear in the experiment can be viewed as one set. One subject from this set of three would be assigned to each of the three groups (e.g., 1, 3, 2 for the first set; 2, 1, 3 for the second set; and so on). Importantly, the group to which any individual is assigned within a set is random. All of this is easily specified on Web-based random numbers generators. In our hypothetical study, we specify that we want 30 sets of numbers and we want each set to include 1, 2, and 3. This will give us an N = 90 (3 groups $\times$ 30 sets), and each group will have an n = 30. Assigning subjects based on numbers drawn in this way ensures that, as the experiment progresses, groups will not differ in size and that subjects in each group are run over the course of the experiment.

Random assignment obviously is important and seems too basic to warrant comment. However, the simplicity of random assignment as a procedure, i.e., how it is accomplished, belies greater nuances. As I discuss later, random assignment does not necessarily guarantee that groups are equivalent. Even so, random assignment can make implausible the likelihood that selection bias (as a threat to internal validity) explains any differences between groups (e.g., experimental and control groups).

Although randomly assigning cases to conditions is the preferred method of assigning subjects, in many situations in which researchers work (e.g., clinics, hospitals, schools), this is not possible. This does not in any way doom the study to weak inferences. Indeed, one's knowledge of principles and practices of methodology becomes more important in this context to ensure that valid and strong inferences can be reached.

There are ways to match subjects between groups when random assignment cannot be accomplished and some of the matching techniques are very sophisticated and can make implausible selection factors as a rival explanation of the results. I will mention those techniques later. In addition to matching, different designs we discuss that are not true-experiments (e.g., quasi-experimental and observational studies) will illustrate ways of drawing strong inferences without the possibility of random assignment.

5.3.2: Group Equivalence

Random assignment is important as a means of distributing characteristics of the sample among groups. There are several subject characteristics (e.g., age, sex, current historical events, motivation for participation), circumstances of participation (e.g., order of appearance or entry into the study), and other factors that might, if uncontrolled, interfere with interpretation of group differences. In some studies, evaluating the impact of these variables may be the central purpose. In other studies, they might be regarded as nuisance variables that, if uncontrolled, will obscure interpretation.

Nuisance variables essentially are those characteristics in which one is not interested but that in principle could influence the results.

In any given study, what counts as a nuisance variable (e.g., some subjects engage in self-injury, others are depressed, and some are tall, others are annoying) could be the main independent variable in another study.

Random assignment is a way of allocating nuisance variables, so they are distributed unsystematically across groups so that the likelihood of selection bias is minimal.

An advantage of random assignment is that it does not require the investigator to be aware of all of the important variables that might be related to the outcome of the experiment. Over a sufficient number of subjects, the many different nuisance variables can be assumed to be distributed evenly among groups.

Random assignment sometimes is viewed as a dependable way of producing equivalent groups. Yet, random assignment refers only to the method of allocating subjects to groups and in a given experiment has no necessary connection with a particular outcome. Randomly assigning subjects can produce groups that differ on all sorts of measures. In fact, we can say more than that. By the very definition of "random," we can expect that groups will be quite different, at least occasionally.

Group differences following random assignment are more likely when sample sizes are small and when there are extreme scores in the sample (Blair, 2004; Hsu, 1989). (When I say group differences, I do not necessarily mean statistically significant differences, but rather genuine differences in characteristics of the sample. The small sample size might not permit us to show that the genuine differences are statistically significant, but that does not mean the groups are no different, as elaborated further below.) As an extreme example, if there are 15 subjects to be allocated to three groups and the subjects vary widely in age, level of anxiety, and other subject variables, it is quite possible that groups may differ on these variables even after random assignment. There are so few subjects that one or two subjects in one of the groups could easily lead to large changes in mean age or level of anxiety. At the other extreme, randomly assigning 150 or 1,500 subjects in the same way would be much less likely for any small set of subjects to make groups depart on some characteristic.

It is important to underscore that random assignment does not necessarily produce equivalent groups. With

random assignment, the likelihood that groups are equivalent increases as a function of the sample size. This means that with small samples group equivalence cannot be assumed. When the total sample (N) is in the range (e.g., 20–40 subjects total in a two-group study), the likelihood that groups are not equivalent across a number of nuisance variables is relatively high (see Hsu, 1989). The net effect is that at the end of the study, the difference between groups due to the experimental manipulation may be obscured or misrepresented because of the nonequivalence of groups. *An effect of experimental manipulation may be diminished or hidden (no statistically significant differences) because of the impact of such variables on outcome.*

Alternatively, some unknown characteristic more evident among subjects in the experimental condition may have led to group differences; it looks as if the experimental manipulation explains the group differences when in fact selection was hidden in the groups. I say "hidden" only to mean not easily detected when the data are analyzed.

The data analysis that we as investigators usually do is intended to establish that the groups are equivalent after random assignment. We compare groups after their random assignment on such variables as age, sex, IQ, years of institutionalization, and pretest performance on the measure of interest. The absence of differences (nonsignificant *t* or *F* tests) may provide false comfort that the groups are equivalent. When the samples are relatively small, statistical power (sensitivity) to detect differences is weak. Thus, the situation in which random assignment is least likely to obtain equivalence (small samples) also is one in which such differences may be the most difficult to detect.

Investigators may feel that the absence of significant differences will satisfy others (e.g., reviewers and advisors) and it usually does. However, the systematic variation that was not detected between groups can still obscure the findings and lead to misleading results. With larger samples, the absence of differences between groups on subject variables and other measures administered before the experimental manipulation provides greater assurance of group equivalence. Even so, such results do not establish absolutely that the groups are equivalent. Groups still may differ on some variable, relevant or irrelevant to the experimental manipulation and performance on the dependent measures that the investigator did not assess.

Random assignment remains vitally important as a concept and procedure. Many statistical tests depend on random assignment. From a methodological standpoint, random assignment makes less plausible or implausible threats to internal validity related to selection. So the benefits of randomization do not require that groups be perfectly equivalent. There is a belief that the procedure guarantees group equivalence in situations when this is not likely, i.e., when the sample size is relatively small. There is no single number one can provide that eliminates the possibility of inequality between or among groups but as a guide >40 subjects per group (*n* not N) is a useful guide for increasing the confidence in the equivalence of groups. As we discuss later, statistical power is the major consideration in deciding the size of the sample one should use. The present discussion focuses attention on a related consideration, namely, more subjects make less plausible selection bias (differences between groups) related to nuisance variables when subjects are assigned randomly to different groups or conditions.

5.3.3: Matching

Often the investigator does not wish to leave to chance the equivalence of groups for a given characteristic of the sample. If a specific subject variable is known to relate to scores on the dependent measure, it is important to take this variable into account to ensure that groups do not differ prior to treatment. For example, it is possible that randomly assigning clients seeking treatment for anxiety could result in one of the treatment groups having participants who were more anxious prior to treatment than those in one of the other groups.

> Group differences after treatment could be directly influenced by the severity of anxiety of the groups before treatment began.

It is undesirable to allow groups to differ prior to the intervention on a variable that is highly related to performance on the dependent measure. The best way to ensure equivalence of groups on a particular dimension is to match subjects on the dimension and then to assign subjects randomly to groups.

> Matching refers to grouping subjects together on the basis of their similarity on a particular characteristic or set of characteristics.

By matching, subjects at each level of the characteristic appear in each group, and the groups will not differ on that characteristic prior to the experiment.

Matching can be accomplished in different ways. Consider, for example, a two-group experiment that is designed to investigate how individuals with depression cope with experimentally induced stress. Prior to the investigation, subjects complete a measure of depression. One way to match subjects is to look for pairs of subjects with *identical scores*. When two subjects are found with the same scores, each is assigned to one of the two groups in an unbiased fashion (e.g., using a random numbers table or coin toss). This is continued with all pairs of subjects with identical scores. If enough pairs of subjects are available and are assigned to groups, mean depression scores for the groups would be identical. Yet, looking for sets of identical scores to match subjects is usually prohibitive because it means

that most subjects who did not have a score identical to another subject's score would not be used. Also, if one wishes to match cases for a three-group (or more group) study, the identical score procedure is virtually possible. There is a better alternative for matching.

A more commonly used procedure is to *rank all of the subjects*, in this case from high to low depression scores. If there are two groups in the experiment, the first two subjects with the highest scores form the first set or block. These two subjects are assigned randomly and individually, so that one member of this set appears in each group. The two subjects with the next highest scores form the next block and are assigned randomly to each group, and so on until all subjects are assigned. This method of assignment utilizes all of the subjects by drawing them from the ranks in blocks of two (or whatever number of groups there are) and assigning them randomly to each of the groups.

> Matching, when followed by random assignment, can equalize groups on the characteristic of interest. The advantage of this procedure is that it does not leave to chance the equivalence of groups on the characteristic(s) of interest.

In some cases, the investigator may wish to ensure that the groups are equivalent on a categorical variable such as subject sex or ethnicity. Random assignment may not ensure that the proportion of subjects assigned to each group will be the same. One way to avoid this problem is to develop the random order of assignment of cases to conditions, as already discussed, but to have separate lists for, say, males and females. If the first two subjects who arrive at the experiment are males, they are assigned (randomly) to each of the two groups (e.g., experimental group, control) of the experiment. If the next person to arrive is a female, she is assigned randomly to the first condition on a separate list for female subjects. Assignments continue in this fashion based on the separate lists. Since each list includes a long stream of 1s and 2s (to indicate assignment to group 1 or 2), the proportion of subjects of each sex will be equal or close to equal no matter how many males or females come into the study. If the overall ratio of males to females who participate in the study is 3:1, this ratio will be reflected in each of the groups. One refers to this in describing the procedure as random assignment with the restriction that an equal number of cases of each sex were assigned to each condition.

Implicit in the discussion is interest in the nature of the variables that are used for purposes of matching. Subjects are matched on these variables that are either known or assumed to be related to performance on the dependent measure. For example, in a study designed to reduce HIV risk behaviors of drug abusing men who engaged in sexual behaviors with other men (a high risk group for HIV), the investigators blocked subjects on a categorical variable (HIV positive or not HIV positive) and assigned randomly to groups so that groups included both types of cases (Mansergh et al., 2010). It is reasonable to consider that these two groups might differ in their likelihood of risky behavior or in their responsiveness to interventions designed to reduce risk. Matching and random assignment removed this variable (HIV status) as a possible source of selection bias.

In general, matching is not essential or inherently valuable in its own right. An investigator matches groups when she knows or suspects that the characteristic relates to performance on the dependent measures. Stated another way some nuisance variables might make an important difference to the conclusions that can be reached. One might match on one or more of those to guarantee rather than hope these variables are distributed across groups. Matching (blocks) and random assignment resolves this issue.

5.3.4: Matching When Random Assignment is Not Possible

The critical component of matching in true-experiments is random assignment. Subjects are matched first and then randomly assigned to groups to distribute across groups the variable on which subjects were matched.

That practice can greatly increase the likelihood that groups are equivalent on a particular variable of interest. Yes, it is possible that some other nonmatched (nuisance) variable still varies across groups, but that is always possible. The concern would be if matching on one variable somehow inadvertently makes the groups unequal on yet another one.

Consider a very different use of matching that is outside of the context of true-experiments. The matching usually occurs in studies where there are intact or pre-formed groups and random assignment is not possible. When subjects are not assigned to groups randomly, the likelihood of subject selection bias (group differences) before any intervention or experimental manipulation is a worry.

> One way to develop groups that are matched is called *propensity score matching*.[4] This is a statistical procedure that integrates multiple variables that may influence selection when groups are compared on a particular outcome. The goal is to devise groups that are matched on all of the variables or at least a large set that contributed to group selection, i.e., those variables for whatever reason led some subjects to be in one condition or group rather than the other group. The feature is that the outcome or the conditions to which participants have been "assigned" (through self-selection or being in a particular setting such as a school or classroom) already have taken place and are not random.

We might, for example, want to know whether eating a vegan diet versus non-vegan diet affects the onset of some disease year later (outcome). We cannot randomly assign individuals to be vegan diet or non-vegan diet types; people self-select for that. And, if we were to find differences in disease outcome, it would be that a host of other variables that covaried were associated with diet. Some of those variables associated with vegan eating might be exercise, alcohol use or abuse, cigarette smoking, parents' eating habits or style of parenting, education, love of methodology, and the list goes on and on of variables that are associated with eating a certain kind of diet. That is, the likelihood of being in the two groups of interest (vegan diet vs. non-vegan diet) is predicted by a long list of other variables. If we want to evaluate diet, we would like to have groups that do not differ on all or most of these other variables. By making groups that are equivalent, we can look at the impact of diet.

> Propensity score matching develops groups that are equivalent by simultaneously matching on multiple variables that could be related to being in the different groups (e.g., vegan vs. non-vegan diet). This is a mathematical solution of integrating multiple variables and estimates the effect of some other variable (e.g., vegan diet, smoking, treatment experience) once these are controlled or integrated in the analyses.

We may not know all the variables in advance that might relate to whether a person is or is not a vegan dieter, but we select and measure multiple variables that might be related.

A summary score is provided that integrates background and potentially confounding variables to provide groups that are equivalent except for the independent variable of interest.

Consider an example. Several years ago, the U.S. government invested millions of dollars annually in programs designed to promote abstinence from sex among adolescents. The goal was to reduce the rates of unwanted pregnancy and sexually transmitted diseases. Individuals were asked to agree to abstinence and take a "virginity pledge." This was a national movement with enormous numbers participating (e.g., >10% of adolescents by mid-1990s). A critical question is whether and to what extent taking the pledge influences actual sexual behavior. Of course, taking or not taking the pledge is not randomly assigned, so we have the concern that any difference (or absence of differences) might be explained by these other variables that make the groups not equivalent. For example, taking the pledge was associated with religious programs, so participation in and commitment to religion is just one of many variables and might contribute to or completely explain the differences between groups in sexual activity.

In a large-scale study using a nationally representative sample, adolescents under 15 years of age who had taken the pledge were compared with other adolescents of the same age who had not taken the pledge (Rosenbaum, 2009). An effort was made to match groups on many variables, actually 112 variables, using propensity score matching. Among the variables were sex (male, female), religion and religious activities, attitudes toward sex, having friends who drink alcohol, parents born in the United States, vocabulary score, and many more. With groups matched and equivalent on many variables that might relate to sexual behavior, now the outcome can be evaluated.

Five years after taking the pledge, individuals who took the pledge and their nonpledged matched comparison peers were surveyed for a variety of sexual activities. Individuals who took the pledge did not differ in rates of premarital sex, sexually transmitted diseases, and anal and oral sex.

Fewer pledgers used birth control during sex than matched nonpledgers.

This latter finding in many ways is the most telling and perhaps disappointing. The goal of the virginity pledge program was to decrease sexually transmitted diseases and unwanted pregnancy. The results indicated that individuals who took the pledge protected themselves *less well* than matched individuals who did not take the pledge. In short, the pledge if anything was associated with a worse outcome. An interesting aside, after 5 years, 82% of those who took the pledge denied ever having taken the pledge.

The example conveys the power of matching apart from the intriguing results. Without being able to assign subjects randomly, the groups were matched on over 100 variables to obtain propensity scores. It is extremely unlikely that there was a selection bias, i.e., differences between groups on variables that could explain the results. Increasingly, propensity score matching is used to evaluate interventions when groups are not comprised randomly as I have noted here (e.g., Eisner, Nagin, Ribeaud, & Malti, 2012; Gunter & Daly, 2012). The strength is in being able to match on a large number of variables that can equalize the groups. Even with a large number of variables (covariates) on which groups are matched, it is always possible that some other variable not assessed that is important differentiates the groups. Yet methodology is always a matter of making potential threats to validity less plausible as explanations of the findings and propensity score matching greatly aids in doing that.

5.3.5: Perspective on Random Assignment and Matching

We have discussed two broad ways of forming groups:

- Randomly assignment of individuals to groups
- Matching

These are not separate necessarily because we discussed matching and then assigning matched sets of individuals randomly to groups. Then in the discussion of propensity score matching, there was no random assignment because the groups were formed already. Yet, propensity matching can be used with random assignment. Random assignment does not ensure equivalent groups, and propensity analysis can even improve on random assignment by following random assignment with propensity matching, a topic to mention but beyond the present scope.

There is a broader point to make. We would like groups to be equivalent in all the variables except the one we are manipulating (true experiment) or studying (observational study).

> There is no guarantee of group equivalence with any single procedure (random assignment, propensity score matching). It is important not to worship one practice as being the answer, because it is not. The goal is always in relation to threats to internal validity and plausible rival hypotheses.

At the end of the study, we would like to be able to say that differences between (among) groups are not likely to be due to history, maturation, statistical regression, and of things, and of course selection bias! We cannot be certain that one or more of these are still having an influence no matter what we do. Yet, we can make the threats implausible as rival explanations of the results. Random assignment with a respectable number in each group (e.g., >40) and propensity analyses and other ways of matching are efforts to do that.

5.4: True-Experimental Designs

5.4 Identify the RAX notation used in illustrating the sequence of events in a research design

Assigning subjects to groups in an unbiased fashion is one of the major defining characteristics of true experiments. Again, by true experiments we are referring to those studies in which the investigator is manipulating conditions, i.e., controls the delivery of the experimental manipulation or intervention and can allocate subjects to these groups in a random fashion. There are several experimental designs. This section discusses different designs commonly used in clinical psychology along with their strengths and weaknesses.

To illustrate the designs, the sequence of events in the design (assessment, intervention) for each group will be presented symbolically using the following notation:

- R stands for Random Assignment of subjects to conditions
- A for Assessment
- X for the Experimental Manipulation or Intervention

The symbols are presented in temporal order so that, for example, A_1 X A_2 signifies that the first observation or pretest (A_1) was followed by an experimental manipulation (X) followed by the second observation or posttest (A_2).

5.5: Pretest–Posttest Control Group Design

5.5 Describe the pretest–posttest control group design

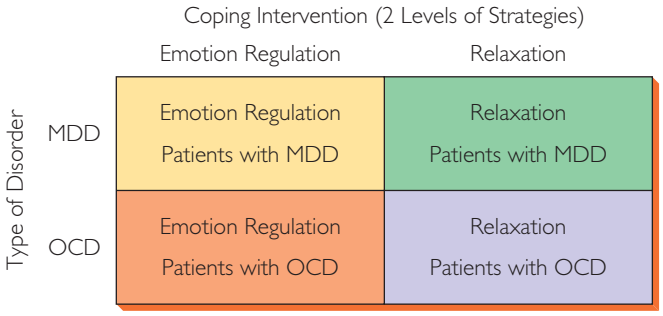
The pretest–posttest design consists of a minimum of two groups. One group receives experimental manipulation or intervention and the other does not. The essential feature of the design is that subjects are tested before and after the intervention, i.e., there is some pretest. Thus, the effect of the manipulation is reflected in the amount of change from pre- to post-assessment.

5.5.1: Description

In the pretest–posttest design, subjects are assigned randomly to groups either prior to or after completion of the pretest. The design can be diagrammed as shown in Figure 5.1.

Figure 5.1: Pretest–Posttest Design

Hypothetical factorial design comparing two independent variables (or factors), Coping Strategy and Psychiatric Disorder. Each factor has two different levels of conditions making this a 2 × 2 factorial design. (Note: MDD stands for Major Depressive Disorder; OCD Obsessive Compulsive Disorder.)



This design applies to any instance in which there is an experimental condition (X) provided to one group and another condition to the other group(s). No "X" between A_1 & A_2 above means that no manipulation or no intervention characterizes the other group. Yet, there can be different control conditions. For example, a study might have one experimental manipulation X_1 and compare that

with another variation of the manipulation X_2. Again, the prototype I provides two groups, but there is no inherent limit to the number of groups, as long as there is random assignment, pre- and post-manipulation assessment, and variation of the experimental manipulation that allows inferences to be drawn about that.

> In many disciplines X is an intervention designed to effect changes in mental or physical health, education, nursing, nutrition, or some other area where there is an applied goal. In such work, the design is called a **randomized controlled trial (RCT)** or randomized controlled clinical trial. This term is a special case of the above that does not affect the design but conveys that the focus is on an intervention (e.g., cognitive therapy, surgery, medication, special educational curriculum).

That is, clients are assigned randomly to receive the intervention and others are assigned to either other interventions or various control conditions, depending on the specific hypotheses. As I mentioned previously, RCTs often are viewed as the "gold standard" for evaluating interventions insofar as many researchers see these as the definitive way of testing an intervention. The gold standard of course is intended to convey that this is the optimal way of establishing the effectiveness of an intervention. The strength and clarity of a pretest–posttest control group design and RCT as a subtype with a special focus are indeed compelling.

5.5.2: An Example of a Randomized Controlled Trial (RCT)

As an example, an RCT was used to evaluate the impact of early intervention for children (ages 2 ½ or under) with autism spectrum disorder (ASD) (Dawson et al., 2010). Children were assigned randomly to receive early intervention program that involved an intensive intervention (2-hour sessions, 2 times per day, 5 days a week for 2 years). The primary intervention was based on applied behavior analysis and focused on developing verbal and nonverbal communication skills in the children. Parents were trained to use strategies they learned in the session at home and for everyday activities (e.g., communication such as play, feeding). Families assigned to the control condition received resource materials and treatment recommendations for other services available in the area, including preschool intervention programs. This might be regarded as a more treatment-as-usual control group insofar as these families received resources often used by families with a child identified with ASD. Assessments were obtained on three occasions (pretreatment, 1 and 2 years after treatment started).

The results were consistent across the 1- and 2-year assessments, so let me note the 2-year assessment to convey the findings. At that latter assessment, the intervention group was significantly better on measures of cognitive functioning and adaptive functioning across multiple domains (e.g., communication, daily living skills, socialization, motor skills). Moreover significantly more children in the intervention group no longer met diagnostic criteria for ASD, compared with children in the control group.

> Children in the control condition actually showed declines rather than improvements in adaptive functioning. The treatment group showed steady improvements in these domains in multiple domains of adaptive functioning.

We can conclude that the intervention program was much more effective than treatment as usual. Random assignment at the beginning of the study followed a matching procedure to equalize IQ, and the proportion of each sex in the groups was matched. Groups were not different at the beginning (baseline assessment). Attrition was not a problem (all cases were retained), and other threats to validity were not plausible. With random assignment and strong differences, this is an optimum and clear test. Early intervention for autism based on applied behavior analysis and an intensive treatment makes a difference and surpasses usual care and use of community resources.

> RCTs are commonly used when there is an interest in seeing if an intervention is effective. When one wants to know whether a particular intervention, program, training regimen of some kind, RCTs usually are used to provide what many believe to be the strongest way to establish that.

In developing evidence-based interventions in a field, including clinical psychology of course, it is the accumulation of RCTs that is recognized as the primary bases for conclusions. The standard is high.

In everyday media blitzes, claims are often made for various psychological interventions, diets, exercise machines, programs to make babies brilliant and strong, and so on. Often terms are used in TV or promotional ads noting that there is "clinical evidence" showing that the program or intervention is effective. Clinical evidence is not a scientific term and not the arbiter of effectiveness. In advertising and marketing, the term "evidence" is not used as it is in science and research methodology. Controlled research studies are used to establish the evidence, and RCTs are a key way of accomplishing that. There are additional strategies to draw causal inferences about intervention effects and these are arguably just as strong, but RCT is considered to be the primary method.

5.5.3: Considerations in Using the Design

Beyond clinical uses of the design (RCTs), the pretest–posttest control group design has several strengths. To begin with, the design controls for the usual threats to

internal validity. If intervening periods between pre- and post-manipulation are the same for each of the groups, threats such as history, maturation, repeated testing, and instrumentation are controlled. Moreover, random assignment from the same population reduces the plausibility that group differences have resulted from either selection bias or differential regression (i.e., to different means). Attrition is not an inherent problem with the design, although as in any experiment that is more than one session differential loss of subjects could interfere with drawing a conclusion about the intervention.

The use of a pretest provides several advantages, as listed in Table 5.1.

Table 5.1: Advantages of Using a Pretest in Research

Advantage	Description
Match subjects	Allows the investigator to match (equalize) subjects on one of the variables assessed at pretest (e.g., level of anxiety) that may influence the results
Evaluate variables	Permits evaluation of that matched variable in the results (e.g., as a separate factor in an analysis of variance or regression analysis)
Statistical power	Increases statistical power of the test
Analyze changes	Allows the investigator to examine who changed, what proportion of individuals changed in a particular way (e.g., show a clinically significant change)
Evaluate attrition	Allows evaluation of attrition (e.g., what were the subjects like who dropped out and did not complete the post-treatment measures?)

First, the data obtained from the pretest allow the investigator to match subjects on different variables and to assign subjects randomly to groups. Matching permits the investigator to equalize groups on pretest performance.

Second and related, the pretest data permit evaluation of the effect of different levels of pretest performance. Within each group, different levels of performance (e.g., high and low) on the pretest can be used as a variable (moderator) in the design to examine whether the intervention varied in impact as a function of the initial standing on the pretested measure.

Third, the use of a pretest affords statistical advantages for the data analysis. By using a pretest, within-group variability is reduced and more powerful statistical tests of the intervention, such as analyses of covariance or repeated measures analyses of variance, are available than if no pretest were used. That is, for a given number of subjects in a study, power is greatly increased if a pretest is used for those subjects than if it is not. This advantage alone is a strong reason to use a pretest because so many studies have insufficient statistical power to detect differences between groups.

Fourth, the pretest allows the researcher to make specific statements about change, such as how many clients improved or actually became worse. In clinical psychology, counseling, education, and rehabilitation where individual performance is very important, the pretest affords information beyond mere group differences at posttreatment. One can evaluate the persons who did or did not change and generate hypotheses about the reasons. The pretest permits identification of the persons who changed or who changed by a specific amount.

Finally, by using a pretest, one can look at attrition in a more analytic fashion than would be the case without a pretest. If subjects are lost over the course of the study, a comparison can be made among groups by looking at pretest scores of those who dropped out versus those who remained in the study. If only a few subjects dropped out, a comparison of dropouts and completers may not be very powerful statistically. Yet, the comparison may show differences, may generate hypotheses about who drops out and why, or may suggest that even with very lenient criteria (e.g., $p < .20$) dropouts and completers do not seem to differ on the variables evaluated. The pretest allows examination of the plausibility of these alternatives.

5.5.4: Additional Consideration Regarding Pretest–Posttest Design

There are some weaknesses to the pretest–posttest treatment control group design. The main restriction pertains to the influence of administering a pretest. A simple effect of testing, i.e., repeatedly administering a test, is controlled in the basic design. What is not controlled is the possibility of an interaction of testing x treatment or a *pretest sensitization effect*. Possibly the intervention had its effect precisely because the pretest sensitized subjects to the intervention.

> A pretest sensitization effect means that the results of the study can be generalized only to subjects who received a pretest.

Whether there is a pretest sensitization effect cannot be assessed in this design. The likelihood of sensitization depends upon several factors. If assessment and the intervention are not close together in time or are unrelated in the perceptions of the subject, sensitization probably is less likely. Therefore, a pretest administered immediately prior to an intervention in the context of the experiment is more likely to lead to sensitization than is assessment in a totally unrelated setting (e.g., in class or in a door-to-door survey at the subject's home) several weeks prior to treatment. Yet the more remote the pretest from the posttest in time and place, the less adequate it may be as a pretest. Intervening events and processes (e.g., history, maturation) between pretest and posttest obscure the effects that can otherwise be more readily attributed to the

experimental manipulation. In general, the strengths of the design clearly outweigh the threat that pretest sensitization will obscure the findings. The information about subject status prior to intervening, the use of this information to match cases and to evaluate change, and the statistical advantages are compelling.

In the context of RCTs, there are additional weaknesses or considerations that can emerge. These are not raised by the design (arrangements of assessment and conditions) but rather the fact that an intervention is being evaluated.

1. In an RCT participants must agree to be assigned randomly to one or more treatments or control conditions. Once assignment is made, participants may drop out immediately because they did not receive the hoped for condition. If more than one treatment is available, clients may remain in the study but not receive the treatment they preferred. That is not trivial because treatment outcomes are better when participants receive their preferred treatment (Swift & Callahan, 2009).
2. Ethical issues are raised in RCTs when an intervention expected to be better is compared to a control condition or no treatment, especially in the context of treating life-threatening conditions (e.g., contracting HIV/AIDs) (Osrin et al., 2009; Solomon, Cavanaugh, & Draine, 2009). Often RCTs are stopped early, before all the subjects are run if it becomes clear that the intervention is having impact and the control condition is not (Bassler et al., 2010). In this way, fewer individuals are exposed to a condition that may not be effective or is much less effective than the other condition in the study.

We will cover ethical issues and protections later on in relation to control and comparison treatment conditions. At this point, it is important to note the special value of the pretest–posttest control group design broadly in addition to the special case of RCTs when interventions are evaluated. Yet, the concerns convey an overarching lesson of methodology that any single method has its weaknesses and reliance on any one study or one method (design, assessment) is risky.[5]

5.6: Posttest-Only Control Group Design

5.6 Contrast the posttest-only control group design with the pretest–posttest control group design

The posttest-only design consists of a minimum of two groups and essentially is the same as the previous design except that no pretest is given.

5.6.1: Description

The effect of the intervention in the posttest-only design is assessed on a post-manipulation measure only. The design can be diagrammed as follows:

$$\begin{array}{ccc} R & X & A_1 \\ R & & A_1 \end{array}$$

(Again, *R denotes that subjects are assigned randomly; A_1 assessment on one occasion, and X the experimental condition*.)

The design controls for the usual threats to internal validity in much the same way as the previous design. The absence of a pretest means that the effect of the manipulation could not result from initial sensitization. Hence, the results could not be restricted in their generality to only those subjects who have received a pretest.

> Often a pretest may not be desirable or feasible. For example, in brief laboratory experiments, the investigator may not wish to know the initial performance level or to expose subjects to the assessment task before they experience the experimental manipulation.

Also, large numbers of subjects might be available and randomly assigned to different conditions in such experiments. With large numbers of subjects and random assignment to the groups, the likelihood of group equivalence is high and the reassurance of a pretest may not be considered worth the effort.

Certainly another feature that must be considered is that a pretest is not always available in clinical research. In many cases, the assessment effort is very costly and a pretest might be prohibitive. For example, an extensive battery of tests might serve as the outcome (posttreatment) measures. The time required to administer and interpret an assessment battery may be several hours, which might make the pretest not worth the cost or effort. In terms of cost, clinical studies frequently rely on neuroimaging techniques, and assessment on multiple occasions (pre and post) easily goes into thousands of dollars for each assessment occasion. In many circumstances from a practical standpoint, there may be no alternative but to omit the pretest. However, a valuable compromise is to administer only one or a few measures from the larger battery at pretest. Ethical considerations also may argue for omission of the pretest, if for example, a pretest might be stressful or invasive (e.g., taking blood samples, asking sensitive personal questions).

5.6.2: Considerations in Using the Design

Understandably, the design is less popular than the one in which a pretest is used. The lack of a pretest raises the discomforting possibility that group differences after the manipulation might be the result of differences between

groups already evident before subjects received their respective conditions. Of course, random assignment of subjects, particularly with large numbers of subjects, is likely to equalize groups. And there is no more likelihood that random assignment will produce different groups prior to the experimental manipulation with this design than in the previous design. Yet, it is reassuring to researchers and consumers of research to see that in fact before the experimental manipulation was presented, the groups were similar on key measures. That assurance can be false (e.g., small sample sizes and no significant difference at pretest do not mean no real difference). Yet it can also be helpful. If there are group differences that difference can be considered in the data analysis to minimize or remove its bias on the results.

> In research with patient populations, the absence of the pretest makes this design less popular for additional reasons. With clinical samples, it is often critical to know the level of functioning of persons prior to the investigation.

For example, a laboratory-based study may compare how individuals with a particular diagnosis will respond to some manipulation. An initial assessment usually is needed to ensure that subjects meet screening criteria (e.g., have experienced some life event in their past) and that control subjects have not. At this time, one might just as well include other measures that will serve as a pretest of the dependent variable in the study. Similarly, an intervention study designed to treat or prevent some dysfunction, the matter of screening individuals prior to the intervention for selection purposes is also relevant. Here too pretest can be readily added to that assessment and gave all of the advantages noted in the previous table. Thus, clinical research tends to favor pretest because some initial assessment is essential anyway for subject selection.

The weaknesses of the posttest-only control groups design derive from the disadvantages of not using a pretest and in any given situation it may not be a "weakness." Thus the inability to ensure that groups are equivalent on the pretest, to match subjects on pretest performance prior to random assignment, or to study the relation between pretest standing and behavior change; the lack of pretest information to evaluate differential attrition across groups; and reduced statistical power are all consequences of foregoing a pretest. Yet, many of these features may not be relevant (e.g., dropping out is not relevant or much of an issue in a one or two-session laboratory study). As one designs a research project, if it possible and feasible to include a pretest, it is advisable to do so in light of the many advantages. Indeed, the increased statistical power alone would make it worthwhile because so many studies in psychological research are underpowered.

5.7: Solomon Four-Group Design

5.7 Analyze the pros and cons of the Solomon four-group design

The effects of pretesting (pretest sensitization) were discussed in each of the above designs. Next we will examine the Solomon four-group design.

5.7.1: Description

The purpose of the Solomon four-group design is to evaluate the effect of pretesting on the effects obtained with a particular intervention (Solomon, 1949).

That is, does administering a pretest in fact influence the results?

At first blush, the design seems highly esoteric—few researchers seemed to have used the design and rarely can one call up an example from memory. (My informal survey at a local supermarket asked 20 people coming through the express cash register line to tell respond to the question "Quick—what is a Solomon-Four Group Design?" Out of 20 people, 19 people did not know; the 1 other person said he thought it was in the produce section.) Actually, there are plenty of examples of the design in contemporary research (e.g., Petersen, Hydeman, & Flowers, 2011; Portzky & van Heeringen, 2006; Rubel et al., 2010). Moreover, interest in and concerns about the effects of pretesting are alive and well and not minor. One reason is that the pretest is so valuable for research for reasons noted earlier in the chapter. Thus, there is a need to see if the pretest in fact contributes to the results. Related, many interventions are designed to address critical health outcomes, and there is interest in many studies in determining whether initial assessment contributes to the effectiveness. That is critical because once an intervention is demonstrated to be effective in controlled laboratory-like conditions, the goal is to extend this more broadly. What if the pretest contributed to the outcome and ended up being pivotal to the effects of the intervention? One would want to know that because merely extending the intervention to community application without the pretest would not be expected to be as effective.

What do we know at this time? Probably the influence of a pretest on sensitizing people to an intervention will depend on the nature of the pretest, the potency of the intervention by itself, the focus of the research, and so on. For example, in the RCT mentioned earlier for the treatment of ASD, preassessments of core cognitive and adaptive skills and symptoms of the disorder are not likely to have influenced the assessments 1 and 2 years later. Many of the symptoms of autism are not nuanced, and changes in these are not likely to respond to sensitization, but I do

not know that to be true and it may be readily argued. We do know two points from current reviews (e.g., Glenn, Bastani, & Maxwell, 2013; McCambridge, Butor-Bhavsar, Witton, & Elbourne, 2011):

1. In the context of interventions for psychological, medical, and school functioning, pretest sensitization can influence the results and yield changes that would not be evident from the intervention alone.
2. Too few studies are done of sufficient quality to clarify the scope and limits of the impact of pretest sensitization.

The research question about sensitization is not about a methodological nuance. We want to make our interventions more effective.

Indeed, most interventions (e.g., information from government agencies, reports, suggestions, news shows) for psychological and physical health (e.g., eat this food, exercise that way) are very weak in terms of their impact. If there were a way to sensitize individuals so these were more effective, that would be a gain. Pretest sensitization might merely be one way but open up a way to increase the effects of otherwise weak interventions.

To address the question of whether there is a pretest sensitization, four groups are required and for ease of reference I have numbered the groups from 1 to 4. These four groups in the design are the two groups mentioned in the pretest–posttest control group design (groups 1 and 2) plus the other two groups of the posttest-only control group design (groups 3 and 4). The Solomon four-group design can be diagrammed as follows. I have changed the numbering of Assessments (A here) for reasons that will be clearer as follows:

1.	R	A_1	X	A_2
2.	*R*		A_3	A_4
3.	R		X	A_5
4.	R			A_6

5.7.2: Considerations in Using the Design

The design controls for the usual threats to internal validity. The effects of testing per se can be evaluated by comparing two control groups that differ only in having received the pretest (i.e., comparison of A_4 and A_6). More important, the interaction of pretesting and the intervention can be assessed by comparing pretested and unpretested groups (i.e., comparison of A_2 and A_5). Actually, the data can be analyzed to evaluate the effects of testing and the testing *x* treatment interaction. To accomplish this, the posttreatment assessment data for each group are combined into a 2 × 2 factorial design and analyzed with a two-way analysis of variance. Only the following observations are used – A_2, A_4, A_5, and A_6. The factors in the analysis are testing (pretest vs. no pretest) and treatment (treatment vs. no treatment). Other methods of analyzing the data from the design are available (see Braver & Braver, 1988; Sawilowsky, Kelley, Blair, & Markman, 1994).

Another feature of the design is that it includes replication of intervention and control conditions. The effect of treatment (X) is replicated in many different places in the design. The effect of an intervention can be attested to by one within-group comparison (A_1 vs. A_2) and several between-group comparisons (e.g., A_2 vs. A_4 or A_6; A_5 vs. A_6 or A_4; A_5 vs. A_3 or A_1).

If a consistent pattern of results emerges from these comparisons, the strength of the demonstration is greatly increased over designs that allow a single comparison.

The design is elegant in the careful way in which pretest sensitization is evaluated. Yet, in methodology as in life most things are trade-offs. In the Solomon-four group design, a great effort goes into evaluating sensitization. Among the trade-offs is statistical power. If one has, let us say, N = 120 subjects for an experiment, putting all of those into four groups would be 30 per group. If group size (n) were increased using the same overall set of subjects (N), statistical power would be increased as well. This can be done with the Kazdin-on-the cheap-three-group version of the Solomon design, which does just that. Use three groups as follows:

1.	R	A_1	X	A_2
2.	R	A_1		A_2
3.	R		X	A_1

This variation evaluates the effects of the intervention (X) with and without the pretest (groups 1 and 3) and keeps in a control group to handle repeated testing (group 2). The 120 subjects now are 40 for three groups—greater power than four-group version. If one is interested in evaluating pretest sensitization, the full four-group version is still the clearest. Yet, the three-group version still provides a test of the impact of the pretest on the outcome by the direct comparison of groups 1 and 3. A rival explanation for any difference there might be that group 1 had the test two times and this is just repeated testing. But that effect is covered in group 2. This Kazdin minimal version is much more feasible to peek at pretest sensitization.

The design may appear to be somewhat esoteric because sensitization effects rarely enter into theoretical accounts of clinical phenomena. Yet, sensitization occasionally has important implications beyond design considerations, as I mentioned. Additional research efforts probably should be directed at studying the effects of pretesting. The pretest–posttest control-group design is used extensively, and the influence of pretesting is rarely studied in the contexts of clinical research. A few studies using

the Solomon four-group design or the Kazdin-on-the-cheap-three-group version in well-researched areas might be very valuable. Demonstrations across dependent measures might establish that in clinical research with widely used measures or interventions, pretest sensitization is restricted to a narrow set of conditions or may not occur at all. As importantly, if there is a way to increase the effectiveness of interventions or experimental manipulations through sensitization experiences of any kind, that would be important to know.

5.8: Factorial Designs

5.8 Express the relevance of the factorial designs when there are multiple variables

The previously mentioned designs consist primarily of evaluating the impact of a single independent variable. For example, the independent variable may be given to one group but withheld from another group. Alternatively, different versions of experimental condition might be provided across several groups. Whatever the variations, the studies basically evaluate one independent variable.

> The main limitation of single-variable experiments is that they often address relatively simple questions about the variable of interest.

The simplicity of the questions should not demean its importance. In relatively new areas of research, the simple questions are the bedrock of subsequent experiments. However, more complex and refined questions can be raised. For example, a single-variable experiment might look at the impact of two different strategies (e.g., emotion regulation, relaxation) for handling experimentally induced stress in a single experimental session. The simple question of which strategy works better is a reasonable focus.

A more nuanced question might be raised by adding a moderator. Perhaps there is reason to believe that the strategies work well with different clinical problems (e.g., depressed vs. obsessive compulsive patients).

> Factorial designs allow the simultaneous investigation of two or more variables (factors) in a single experiment. Within each variable, two or more levels or conditions are administered.

In our hypothetical example, we have two variables:

1. Type of coping strategy
2. Type of clinical problem

Each variable has two levels (regulation or relaxation for the coping variable; depression or obsessive compulsive disorder for the clinical problem variable). This 2 × 2 design (2 variables each with 2 levels) forms four groups that represent each possible combination of the levels of the two factors, as shown in Figure 5.2. The data analyses will identify whether the coping strategies differ from each other on some measure of stress, whether the two diagnostic groups differ, and whether the effects of coping vary as a function of (are moderated by) diagnostic groups.

Figure 5.2: Coping Intervention (2 Levels or strategies)

Hypothetical factorial design comparing two independent variables (or factors), Coping Strategy and Psychiatric Disorder.

	Emotion Regulation	Relaxation
MDD	Emotion Regulation	Relaxation
Type of Disorder	Patients with MDD	Patients with MDD
OCD	Emotion Regulation	Relaxation
	Patients with OCD	Patients with OCD

A major reason for completing a factorial experiment is that the combined effect of two or more variables may be of interest, i.e., their interaction.

> An *interaction* means that the effect of one of the variables (e.g., coping strategy) depends on the level of one of the other variables.

Earlier we discussed interactions in terms of external validity. In this light, the interaction means that the effect of one variable may or may not be generalized across all conditions. Rather, the impact of that variable occurs only under certain conditions or operates differently under those conditions (e.g., with men rather than women, with younger rather than older persons). We also discussed this as moderation, i.e., a variable that influences the magnitude or direction of the relation of two other variables.

Each factor has two different levels of conditions making this a 2 × 2 factorial design. (Note: MDD, Major Depressive Disorder; OCD, Obsessive Compulsive Disorder.)

A factorial design is not a single design but rather a family of designs that vary in the number and types of variables and the number of levels within each variable. The variation of factorial designs also is influenced by whether or not a pretest is used. If a pretest is used, testing can become one of the variables or factors (time of assessment) with two (pretest vs. posttest) or more levels. The data can be analyzed to assess whether subjects changed with repeated assessment, independently of a particular intervention.

In single-variable experiments, one manipulation is of interest and all other variables that might influence the results are controlled. In a factorial experiment, multiple variables are included to address questions about separate and combined effects of different variables. The variables

that are included in the factorial design are not merely controlled; their effect is evaluated as distinct variables in the design.

5.8.1: Considerations in Using the Design

The strength of a factorial design is that it can assess the effects of separate variables in a single experiment. The feature includes one of economy because different variables can be studied with fewer subjects and observations in a factorial design than in separate experiments for the single-variable study of each of the variables, one at a time. In addition, the factorial design provides unique information about the combined effects of the independent variables.

> The importance of evaluating interactions cannot be overestimated in conducting research.

Essentially, interactions provide the boundary conditions of independent variables and their effects (referred to as generality of the effect) and moderators (other variables that may influence the relation).

The concerns about using the factorial designs are both practical and interpretive. On the practical side, one must remember that the number of groups in the investigation multiplies quickly as new factors or new levels of a given factor are added. For example, a design in its conceptual stages might simply begin as a 2×3 by looking at type of treatment (mindfulness training vs. biofeedback) and severity of anxiety (high, moderate, and low). This design already includes 6 (i.e., 2×3) groups. Yet it also might be interesting to study whether the treatment is administered by a live therapist or prerecorded modules administered by computer. This third variable, manner of administering treatment, includes two levels, so the overall design now is a $2 \times 2 \times 3$ and has 12 groups. Also, while we are at it, perhaps we could explore the fourth variable, instructions to the subjects. This variable might have two levels in which half of the subjects are told that the treatment was "discovered by a reality show survivor" and the other half that it was "discovered by a scientist engaged in basic laboratory research." We might expect mindfulness subjects who receive the "guru instructions" and biofeedback subjects who receive the "scientific researcher instructions" to do better than their counterparts. Now we have a $2 \times 2 \times 2 \times 3$ design or 24 groups, a formidable doctoral dissertation to say the least. Instead of a study, we have a career. As a general point, the number of groups in a study may quickly become prohibitive as factors and levels are increased. This means that the demand for subjects to complete each of the combinations of the variables will increase as well.

> In practice, there are constraints in the number of subjects that can be run in a given study and the number of factors (variables) that can be easily studied.

A related problem is interpreting the results of multiple factor experiments. Factorial designs are optimally informative when an investigator predicts an interactive relationship among two or more variables. Simple interactions involving two or three variables often are relatively straightforward to interpret. However, when multiple variables interact, the investigator may be at a loss to describe the complex relationship in a coherent fashion, let alone offer an informed or theoretically plausible explanation. A factorial design is useful for evaluating the separate and combined effects of variables of interest when these variables are conceptually related and predicted to generate interactive effects. The inclusion of factors in the design is dictated by conceptual considerations of those variables and the interpretability of the predicted relations.

I have mentioned factorial designs to encourage attention to interaction terms, i.e., the combined relation of two or more variables. There are other ways to accomplish this goal. Regression analysis includes a family of ways of analyzing data where multiple variables can be combined to predict a particular outcome. These matters are beyond the scope of this text but available in many statistics texts.

5.9: Quasi-Experimental Designs

5.9 Recognize the areas where the researcher has no control over the subjects as quasi-experimental designs

The previous designs constitute basic between-group experimental designs and are true experiments because central features of the study can be well controlled to eliminate or make very implausible threats to internal validity.

The main feature is the investigator's ability to assign subjects randomly to conditions.

There are many situations in which the investigator cannot exert such control over subject assignment, but the investigator still wishes to evaluate an intervention of some kind. In clinical, counseling, educational research, or more generally research in many applied settings, investigators cannot shuffle participants, clients, or students to use random assignment. There are intact groups or groups in various settings, and one must work within administrative, bureaucratic, and occasionally even anti-research constraints. The investigator may be able to control delivery of the intervention to some clients but not to others but the groups are not randomly composed.

> As noted earlier, research designs in which the investigator cannot exert control required of true experiments have been referred to as quasi-experimental designs (Campbell & Stanley, 1963).

For investigators who are genuinely bothered by less well-controlled studies, these can also be called *quasi-experimental designs*. No matter what they are called, very strong inferences can be drawn from quasi-experimental designs. However, the designs often require greater ingenuity in selecting controls or analyzing the data to make implausible various threats to validity (especially, selection *x* history or selection *x* maturation). As you recall selection in combination with another threat (e.g., selection *x* history) means that the groups might systematically vary in exposure to events (e.g., in the school neighborhood). In a true experiment, these experiences are varied but unsystematic across groups due to random assignment.

5.10: Variations: Briefly Noted

5.10 Examine the nonequivalent control group designs

There are many between-group quasi-experimental designs because various groups might be added to for varied control purposes, the most common of which parallel the pretest–posttest and posttest-only experimental designs. For each of the quasi-experimental equivalents of these designs, the control group is not demonstrably equivalent to the experimental group, usually because subjects have been assigned to groups prior to the inception of the investigation. Because the groups are already formed, they may differ in advance of the intervention. This explains why the designs have also been referred to as *nonequivalent control group designs* (Campbell & Stanley, 1963).

5.10.1: Pretest–Posttest Design

The most widely used version of a nonequivalent control group design is the one that resembles the pretest–posttest control group design. The design may be diagrammed as follows:

nonR	A_1	X	A_2
nonR	A_1		A_2

In this version, nonrandomly assigned subjects (e.g., subjects who already may be in separate clinics, schools, or classrooms) are compared.

One group receives the intervention and the other does not. The strength of the design depends directly upon the similarity of the experimental and control groups.

The investigator must ask how the assignment of subjects to groups originally might have led to systematic differences in advance of the intervention. For example, two high schools might be used to evaluate a drug-abuse prevention intervention in which the intervention is provided at one school but not at the other. Youths in the schools may vary on such factors as socioeconomic status, IQ, or any number of other measures. Possibly, initial differences on the pretest measures or different characteristics of the groups, whether or not they are revealed on the pretest, account for the findings. The similarity of youths across schools can be attested to partially on the basis of pretest scores as well as on various subject variables. Pretest equivalence on a measure does not mean that the groups are comparable in all dimensions relevant to the intervention, but it increases the confidence one might place in this assumption.

In the version of the design diagrammed previously, the results could not easily be attributed to history, maturation, testing, regression, mortality, and similar factors that might occur across both groups. However, it is possible that these threats might differ *between* groups (i.e., selection *x* history or selection *x* maturation). These interactions mean that particular confounding events may affect one group but not the other, and hence might account for group differences. For example, one group might experience historical events (within the school) or differ in rate of maturation (improvements without treatment). These influences might account for group differences even if the subjects were equivalent on a pretest.

Among the options that can be used to reduce the prospect of differences between groups due to confounding factors (also here referred to as covariates), propensity score matching can be used, as highlighted previously. Thus, even though the group assignment is fixed (e.g., to different schools) individuals who receive and do not receive the intervention could readily be matched and further reduce threats to internal validity as plausibility of rival interpretation of the results. Thus, the design can yield strong inferences based on what the investigator does to make implausible those threats that random assignment normally handles, as illustrated further in an example later.

5.10.2: Posttest-Only Design

A nonequivalent control group design need not use a pretest. The posttest-only quasi-experimental design can be diagrammed as follows:

nonR	X	A1
nonR		A1

Of course the problem with this design, as with its true-experimental counterpart, is that the equivalence of groups prior to the intervention cannot be assessed. In the posttest-only experimental design, discussed earlier, the absence of a pretest was not necessarily problematic because random assignment increases the likelihood of group equivalence, particularly for large sample sizes. However, in a posttest-only quasi-experiment, the groups may be very different across several dimensions prior to

the experimental manipulation. Hence attributing group differences to the intervention may be especially weak. Aside from problems of probable group nonequivalence prior to the experimental manipulation and the absence of a pretest to estimate group differences, this version of the nonequivalent control group design suffers from each of the possible threats to internal validity of the same design with a pretest. The absence of pretest information means that one cannot match (propensity scores) groups as one strategy to reduce the plausibility of various threats to validity associated with selection.

At first blush, one might wonder why even to include this design here. The design is still quite useful. It might be especially useful as a preliminary study to see if a program is working or shows promise. It is important to begin with recognition of a lamentable fact that in psychology, education, criminology, rehabilitation, and so on, the vast majority of programs, interventions, and curricula are well intended but not evaluated in anyway. A quasi-experiment and with a posttest only design would be a huge leap in suggesting whether the intervention has promise. Although the posttest-only quasi-experiment is weak, occasionally this may be the most viable design available.

5.11: Illustration

5.11 Illustrate how a quasi-experimental design was used to study the impact of secondhand cigarette smoke

The quasi-experimental designs were briefly covered because they resemble the true-experimental designs that were detailed previously. Yet, the richness of quasi-experimental designs and the methodological thinking behind them is lost with mere presentation of symbols to reflect group conditions and assessment. Consider a between-group study that is a quasi-experiment. There were not quite perfect controls and people were not assigned randomly to groups. This is a study that focused on the impact of secondhand cigarette smoke, which is known to have adverse effects including heart disease. Eliminating smoking in indoor spaces is the best way to protect nonsmokers.

Some cities have instituted smoke-free ordinances that ban smoking in public places (e.g., restaurants, taverns) and work places.

Do you think such ordinances make a difference?

Arguably the best way to test this would be to randomly select cities in the country and then randomly assign a subset of these to be smoke free and others not to be smoke free. This RCT is not going to happen for a host of reasons. Try telling mayors of several cities or governors of various states in the United States that they were assigned to the control condition and everyone in indoor spaces cannot smoke to his or her heart's content (or disease).

If no RCT is possible, what can one do?

A quasi-experiment with the idea of making implausible the threats validity is a good answer.

In one such quasi-experiment (referred to as The Pueblo Heart Study) with several reports, the question has been examined by selecting and comparing three cities (Centers for Disease Control and Prevention, 2009) in a pretest–posttest quasi-experimental design. Pueblo, Colorado, had a smoke-free ordinance and was compared to two nearby cities over a 3-year period. The two nearby cities did not have smoke-free ordinances and served as comparison cities.

The results: In Pueblo, with implementation of its smoke-free ordinance, hospitalization rates for acute myocardial infarction (heart attacks) markedly decreased from before to after the ordinance was implemented. No changes in hospitalization rates were evident in the two comparison cities.

Does this finding establish and prove that secondhand smoking leads to increased heart attack? No, but no one study rarely does that anyway. Also, we would want to know more about the comparability of the three cities and their hospitals, demographic composition of the cities, and more. It is possible that selection or different historical events associated with the cities (selection x history) could explain the findings. Also, was it reduced secondary smoking or more people just quitting smoking, which also results from a ban? All these and more are good questions, but one should not lose sight of the strength of the evaluation. The findings suggest that bans do make a difference. Of course, it must be replicated. It has been. The findings hold. With replication also in a quasi-experiment, threats to validity (e.g., history, maturation, retesting) are not very plausible. Still we need to learn more about what facets of smoking changed and what their specific impact was.

This is a good example because the question was one of huge importance (public health, heart disease, death).

Is there any impact of a public ordinance? It is important to evaluate because if it is effective, we would want to extend this to other cities that were willing to use this approach. If the ordinance is ineffective, what a waste of time and resources! We would want to find that out right away.

There are many situations in which we believe we are helping or we have an idea that we think will make an important difference in society. The challenge is to add evaluation to that. If the most rigorous research can be done, yes always, we seize that opportunity. But the other side is the problem. When the most rigorous study cannot be done, this is not the time to go by our anecdotal experience. Many threats to validity can be made implausible to

help draw valid inferences. This is methodology at its best (using ingenuity to improve the inferences that can be drawn) not methodology at its easiest (random assignment and careful control).

Among methodological purists, occasionally there is the view that an RCT is not only the best way to demonstrate the effectiveness of an intervention but the only way. That is arguable and lamentable. I say lamentable because we know that well-intended programs (e.g., for suicide prevention, unprotected sex, treatment of aggression) occasionally can harm, i.e., they make the target problem worse. This makes evaluation central, whether or not an RCT can be used. Quasi-experimental arrangements are essential in such situations. It is in these difficult-to-evaluate-situations that knowledge of methodology (e.g., threats to validity and how to control them) and various statistical tools (e.g., matching) are utilized the most.

5.11.1: General Comments

Although the nonequivalent control group designs already mentioned constitute the most frequently used variations, all of the possible quasi-experimental designs cannot be enumerated. In other variants, a special control group may be added to address a particular threat or set of threats to validity.

> The additional control group is sometimes referred to as a "patched up" control group to convey that it is a complete add-on but added strategically to address a potential threat.

The general characteristic of quasi designs is that constraints inherent in the situation restrict the investigator from meeting the requirements of true experiments. In such cases, ingenuity is required to mobilize methodological weapons against ambiguity. Various control groups can be used to weaken one or more threats to internal validity and patchup an otherwise imperfect design. Also, matching techniques I have mentioned can help with interpretation of the results.

5.12: Multiple-Treatment Designs

5.12 Recognize crossover design as a form of multiple-treatment design

The defining characteristic of the multiple-treatment design is that each of the different conditions (treatments) under investigation is presented to each subject.

> "Treatment" is used to refer to the design because of the frequent use in the context of evaluating interventions that produce therapeutic change (e.g., psychological intervention, medication).

Yet, it is better to consider treatment here to stand for "condition" and that might be two or more experimental manipulations with or without control conditions. Thus a laboratory study might well present different conditions (e.g., such as exposure to different tasks, different priming experiences, different confederates acting in a particular way).

Although the evaluation of treatments is "within subjects," separate groups of subjects are present in the design. In multiple-treatment designs in clinical research, separate groups are used with the goal of balancing the order of the treatments. Balancing means that different orders are presented so that the effect of the treatment is not confounded by the position (always presented first) in which it appeared. Because separate groups are used in the multiple-treatment designs, points raised about random assignment and matching are relevant for constructing different groups for multiple-treatment designs.

There are different versions of multiple-treatment designs that depend upon the number of treatments and the manner in which they are presented.

All of the designs might be called *counterbalanced designs* because they try to balance the order of treatment across subjects.

However, it is worth distinguishing both the commonly used version of the multiple-treatment design and the general method for balancing treatments.

5.12.1: Crossover Design

A specific multiple-treatment design that is used most often is referred to as the *crossover design*. The design receives its name because part way through the experiment, usually at the midpoint, all subjects "cross over" (i.e., are switched) to the other experimental condition. The design is used with two different conditions. Two groups of subjects are constructed through random assignment. The groups differ only in the order in which they receive the two treatments. The design can be diagrammed as follows:

R A1X1 A2 X2 A3 R A1 X2 A2 X1 A3

The diagram may appear complex because of the numbering of different interventions (X_1 and X_2) and the different observations (A_1 and A_2). However, the design is relatively straightforward. Essentially, each group is formed through random assignment (R). A pretest may be provided to assess performance prior to any intervention. The pretest (designated in the diagram as A_1) is not mandatory but is included because it is commonly used and provides the benefits, discussed earlier. The crucial feature of the design is that the groups receive the interventions (X_1 and X_2) in a different order. Moreover, the subjects are assessed after each intervention. Thus, there

is an assessment halfway through the study at the crossover point as well as after the second and final treatment is terminated.

The balancing of treatment is straightforward. Because there are only two treatments (X_1 & X_2), all possible orders are easy. Here, balancing only means one group receives X_1 and then X_2 and the other group receives X_2 and then X_1. Each treatment appeared in each position (first, second), and each treatment preceded and followed the other. Balancing becomes much more intricate once one adds more treatments but the crossover version includes just two.

> The design is used frequently in evaluating the effects of medication on various symptoms or disorders. For two (or more) medications, a comparison can be made within the same patients if there is an intervening "washout" period during which all medication is stopped and hopefully leaves (is washed out) a person's system. The second medication can then be administered with little or no concern over lingering effects of the first medication.

In psychological experiments, two or more treatments can be provided this way too but it is difficult to continue to show increments of change as one treatment builds on another on outcome measures, for reasons discussed later. Also, it is more difficult to "washout" psychological interventions, i.e., remove completely their prior impact, and it is not really clear what that would mean and how one would show that.

The crossover design is nicely illustrated in a comparison of two conditions (caffeine, placebo) provided to each subject (Smith, Lawrence, Diukova, Wise, & Rogers, 2012). Well known is the influence of caffeine as a stimulant. There are reasons to believe both from human and nonhuman animal studies that caffeine can increase anxiety. This is based on the effects of caffeine on neurons in the brain and how these influence response to external stimuli. The investigators tested the hypothesis that caffeine would increase responsiveness (reaction) to threat and potentially anxiety-provoking cues and these would be reflected in self-report, blood pressure, and activation (fMRI) of brain centers associated with anxiety (e.g., amygdala, especially the basolateral complex). On two occasions, healthy volunteers were exposed to a capsule (pill) that included a caffeinated powder or a placebo powder (of cornstarch). This was a double-blind study so individuals administering the tasks or drink and the subjects could not tell from the capsules what condition was being administered.

Participants came to the experiment on two occasions (one week apart); on each occasion, they were exposed to a task (responding to faces with emotional expressions—happy, sad, angry, fearful, and others) while being in the magnet to evaluate brain activation. As in a crossover design, some subjects were assigned to caffeine first and then placebo and other subjects were assigned to placebo first. Thus the different conditions were balanced, i.e., equally dispersed across subjects and each appeared before and after the other. The caffeine was a single dose of 250mg (equivalent to 2–2½ cups of ground coffee).

> The results supported the prediction. When under the influence of caffeine, subjects showed significantly greater anxiety on the measures, including activation of brain regions associated with processing fear and anxiety. The conclusion was that caffeine can indeed induce heightened response to social cues of threat and anxiety.

The effects were nicely demonstrated in a crossover design, in this case with an experimental and a control condition.

5.12.2: Multiple-Treatment Counterbalanced Design

The crossover design as discussed here is a simple design, usually with two conditions, in which each client receives the different condition but in a different order; that is, the conditions are counterbalanced. With an increase in the number of conditions, however, counterbalancing becomes more complex and the order in which the treatments are given is more difficult to balance. Consider three (A, B, C) rather than two conditions. The conditions require all of the sequences to be completely balanced – ABC, ACB, BCA, BAC, CAB, CBA. These six sequences reflect the order in which the conditions appear. The subjects who are to serve are randomly assigned to one of these groups or sequences in which all three conditions are presented. The design would be completely balanced insofar all possible orders of the treatment (all permutations of ABC in their varying orders) are provided. In practice, this is way too cumbersome (i.e., having all six groups) and the variations increase exorbitantly if one includes four or more conditions.

In practice, a special arrangement is used referred to as a Latin Square. In a Latin Square, each condition (A, B, or C) occurs once and only once in each position. Table 5.2 provides a hypothetical example of three conditions (ABC) presented to subjects. There are three groups of subjects, and each group receives all conditions but in a different disorder. If one looks at the rows (horizontal), it is clear that each condition (e.g., A) appears once and only once in each position (first, second, or third). If one looks at the columns, each condition also appears once and only once in each position. At the end of the investigation, analyses can compare different conditions and can assess whether there were any effects due to groups (rows), order (columns), or condition (As vs. Bs vs. Cs).[6]

Table 5.2: Hypothetical Multiple Treatment Study with Three Conditions

	Position 1	Position 2	Position 3	Mean or Sum of Sequences
Sequence (Group) 1	A	B	C	
Sequence (Group) 2	B	C	A	
Sequence (Group) 3	C	A	B	
Mean or Sum of Rows				

- Order effects refer to a comparison of the columns (means or sums of Positions 1, 2, and 3)
- Sequence effect refers to a comparison of the rows (means or sums of Groups 1, 2, and 3).

For each of these, totals are evaluated that ignore the individual treatments.

One effect not completely controlled is the sequence in which the treatment appears. The sequence of treatments in the table (the rows) does not represent all possible sequences. Not every treatment is preceded and followed by every other treatment. For example, A never follows B, C never follows A, and so on. Hence, it is not really possible with the above design to rule out the influence of different sequences as a contributor to the data for a given condition. There may be an interaction between the effects of treatment and when treatment appears in the sequence. This interaction can be avoided as a source of confound by using all possible orders of treatment with separate groups of subjects. In a completely balanced design, each treatment occurs equally often in each order and each treatment precedes and follows all others. The problem with such a design is that the number of groups and subjects required may be prohibitive. (The number of subjects for complete counterbalancing would be *k* factorial, where *k* equals the number of conditions in the experiment.) Which sequences are selected for a given study is based on random selection from a table of Latin Squares (see footnote 6, noted previously).

In general, the administration of three or more treatments to the same subject is uncommon. When treatment studies use multiple-treatment designs, two treatments are more commonly compared, as illustrated with the crossover design. Conducting additional treatments may require a relatively long period of continuous treatment so that each treatment has an opportunity to influence behavior. Moreover, the problem of reflecting change with multiple treatments, discussed below, makes testing for the effect of several treatments a dubious venture. Consequently several treatments are evaluated within subjects infrequently; when they are, the designs usually are not completely balanced to include all possible sequences of treatment. I mention Latin Square briefly to be familiar when you read or hear it, but most probably researchers do not do a Latin Square study in their careers and most readers may not see one in the articles they read.

5.13: Considerations in Using the Designs

5.13 Identify some of the deliberations that need to be taken into account while choosing a multiple-treatment design

Multiple-treatment designs are used frequently in psychology, medicine, nutrition, and other areas and in both basic and applied research. In basic research, the designs provide opportunities to evaluate procedures or interventions (e.g., different reinforcement schedules, diets, medications, or coping strategies) on immediate outcomes (e.g., rate of lever pressing, indices of metabolism, side effects, neuroimaging) to understand processes under controlled conditions. In treatment with clearly applied goals, the designs may be used as well (e.g., to decide which among two alternatives is more likely to be effective). The utility of multiple-treatment designs depends upon several factors, including the anticipated effects of juxtaposing different treatments, the type of independent and dependent variables that are studied, and the measurement of cumulative treatment effects with the same subjects.

5.13.1: Order and Sequence Effects

Perhaps the most important consideration in using a multiple-treatment design relates to the problem of ordering treatments. Actually there are different problems that can be distinguished. To begin with, if an experiment consisted of only one group of subjects that received two different conditions (A and B) in the same order, the results would be completely uninterpretable. For example, if condition B led to greater change than condition A, it would be impossible to determine whether B was more effective because of its unique properties or because it was the second treatment provided to all subjects. Treatment B may have been more effective because a continuation of treatment, independently of what the treatment was, may have led to greater change. Thus, the order in which the treatments appeared in this single group study might have been

responsible for treatment differences and hence serves as a plausible alternative explanation of the results.

> When the order of treatments might account for the results, this is referred to as an order effect.

The effect merely refers to the fact that the point in time in which treatment occurred, rather than the specific treatment, might be responsible for the pattern of results. In most multiple-treatment designs, order effects are not confounded with treatments because of counterbalancing, as illustrated in the discussion of crossover and Latin Square designs. Although order is not confounded with treatment where counterbalancing is used, it still may influence the pattern of results. For example, treatments presented first may be more effective no matter which treatment it is. Quite possibly the reason for this is related to ceiling and floor effects, discussed later, in that by the time the final treatment is provided in a series of treatments, the amount of change that can be reflected on the dependent measures is small.

There is another way that the specific order of treatments may influence the results. Specifically, the transfer from one treatment to another may not be the same for each treatment. Receiving treatment A followed by treatment B may not be the same as receiving treatment B followed by treatment A. The order in which these appear may partially dictate the effects of each treatment.

> When the arrangement of treatments contributes to their effects, this is referred to as sequence effects.

The nature of the problem is conveyed by other terms that are sometimes used, such as *multiple-treatment interference* or *carryover effects*. The importance of the sequence in which different events appear in dictating their effects is obvious from examples of everyday experience. For example, the taste of a given food depends not only on the specific properties of the food but also upon what food or liquid has immediately preceded it.

Order and sequence effects are potentially confusing. Return to Table 5.2 to see the comparison of three hypothetical conditions (ABC). Ignore the ABCs this time and look to the rows (horizontal) and columns (vertical). Order effect refers to a comparison of Positions 1, 2, and 3 columns. If we sum ABC in Position 1 (which is the first treatment provided to each group, respectively), we can compare that with the sum of Position 2 and the sum of Position 3 treatments are there differences. This ignores what the ABC treatments are and looks at column totals. Sequence effect looks at the sum of the rows. Is Sequence 1 total across ABC treatments any different from the sum of Sequence 2 and sum of Sequence 3? Any differences on those row totals would be a sequence effect.

As a general statement, multiple-treatment designs are quite susceptible to the influence of sequence effects. Whether these effects are viewed as nuisances depends upon the purposes of the investigator.

> Sequence effects represent complex interactions (e.g., treatment *x* order of appearance) and may be of interest in their own right.

All events in one's life occur in the context of other events. Hence, sequence effects embrace questions about the context in which events occur and the effects of prior experience on subsequent performance. Yet, the complexities are in those cases in which three or more conditions are provided to the same subject. If only two (A,B) conditions are presented, there still may be sequence (and order) effects because going from A to B may have different effects from going from B to A. Yet, all is more easily evaluated in the crossover design with only two conditions.

5.13.2: Restrictions with Various Independent and Dependent Variables

Considerations pertaining to the variables that are to be studied may dictate whether a multiple-treatment design is likely to be appropriate or useful for the experiment in question. Certain variables of interest to the investigator are not easily studied in a multiple-treatment design. For example, the experimental instructions, subject expectancies, or a stress induction manipulation may present particular problems, depending upon the precise experimental manipulations. The issue is that there might well be a lingering influence of the first condition that is carried over or is in conflict in some way with the second condition and merely balancing these by crossover (A,B for one group and B,A for the other) is not necessarily helpful. One might not be able to present that repeatedly because the impact might be expected to carry over (e.g., better adaptation to the second stressor). Alternatively, one may use a washout period, which refers to an interval (e.g., 1 week) that is designed to eliminate or reduce any immediate carryover effects. Individuals receive the separate conditions, but some time (e.g., hours, days, weeks) is interspersed with the expectation that time will reduce carryover.

Discussing potentially conflicting interventions or carryover from one condition to the next raises another side of the issue. It is possible to select conditions that are very similar. For example, the "different conditions" presented to the subjects may only vary in subtle characteristics. These "different" conditions may produce few detectable effects in a multiple-treatment design because subjects do not distinguish the conditions:

- The first condition may lead to a certain degree of change.

- The second condition, perhaps just a small variation, may not be perceived as any different from the first one and hence may produce no differences within subjects.

Essentially, the second condition is perceived as a continuation of the first. Although condition differences would not be revealed by changes within subjects, a comparison between groups for the first condition administered might yield a difference. That comparison reflects the impact with no other condition was provided other than the first.

> Personality, demographic, physical, and other stable characteristics are not studied within subjects because they do not vary within the same subject for a given experiment.

Obviously, participants are not both male and female or a psychiatric patient and not a patient within the same experiment. However, it is possible to provide experiences within the experiment (e.g., instructions, expectancies, incentives to perform in one way rather than another) that changes how a subject reacts to certain variables. A participant could be given a success or failure experience in an attempt to assess the impact of these experiences on dependent measures. Stable subject characteristics can be readily studied in factorial designs that combine group and multiple-treatment features. For example, a subject can be classified by one variable (e.g., sex, age, level of anxiety) and receive each of the different levels of another variable (e.g., mood inductions to be happy and sad). This combined design can examine whether mood reactions differ according to subject characteristics.

Aside from restrictions on independent variables, there are restrictions on dependent measures that can be readily evaluated in a multiple-treatment design. Dependent measures involving such skills as cognitive or motor abilities may not readily reflect treatment effects within subjects. When one treatment alters a skill (e.g., bicycle riding or reading), the effects of other treatments are more difficult to evaluate than when transient changes in performance are made.

5.13.3: Ceiling and Floor Effects

A possible problem in evaluating different experimental conditions within the same subjects is that ceiling or floor effects may limit the amount of change that can be shown.

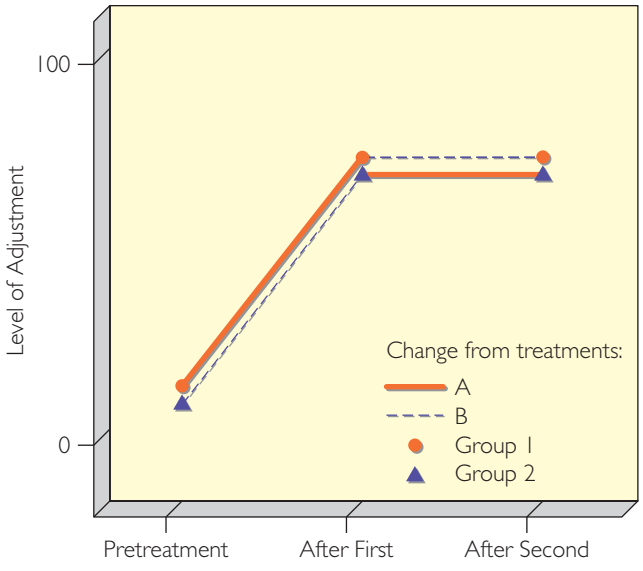
> Ceiling and floor effects refer to the fact that change in the dependent measures may reach an upper or lower limit, respectively, and that further change cannot be demonstrated because of this limit.

The amount of change produced by the first intervention may not allow additional change to occur.

Assume, for example, that two treatments are presented in a multiple-treatment design and evaluated on a hypothetical measure of adjustment that ranges in scores from 0 to 100. Here, a score of 0 equals "poor adjustment," which means the individual is constantly depressed, anxious, drunk, suicidal, and apathetic—and this is on the good days. Assume that 100 equals the paragon of adjustment or that the individual is perfectly adaptive, content, and self-actualizing even in the face of recent loss of family, possessions, job, fortune, memory, and favorite methodology textbook. In pretreatment assessment, subjects are screened and selected based on their poor adjustment on the scale; say, scores lower than 25. Then two treatments are provided, in counterbalanced order, to two groups of subjects. Suppose the initial treatment increases adjustment to a mean of 95. With this initial change, a second treatment cannot provide evidence of further improvements. For example, the data might show the pattern illustrated in Figure 5.3, in which it can be seen that the first treatment (A or B) led to marked increments in adjustment and administering the second treatment did not produce additional change. The conclusion would be that the treatments are equally effective and that one does not add to the other.

Figure 5.3: Hypothetical Data for a Crossover Design

Hypothetical data for a crossover design where each group of subjects receives treatments but in a counterbalanced order.



5.13.4: Additional Considerations Regarding Ceiling and Floor Effects

A different pattern might emerge if there were no restricted ceiling on the measure. That is, if even higher scores were allowed and a greater amount of change could be shown,

different conclusions might be reached. For example, if the adjustment scale allowed scores beyond 100 and additional degrees of adjustment, different results might have been obtained. The treatments might have been different at their first presentation. Treatment A might have led to a mean score of 95 but treatment B to a score of 150. In that case, when the other (second) treatment was applied to each group, additional changes may have been detected, at least in going from A to B.

> In general, the problem of ceiling or floor effects is not restricted to multiple-treatment comparisons. The absence of differences between groups on a measure may result from limits in the range of scores obtained on that measure.

If scores for the different groups congregate at the upper and lower ends of the scale, it is possible that differences would be evident if the scale permitted a greater spread of scores. For example, in child treatment outcome studies of my group, we evaluate treatment acceptability, i.e., the extent to which treatment was viewed as appropriate, fair, and reasonable. At the end of a study, parents and children rated the treatment they received. We have used different evidenced-based treatments (e.g., parent management training, cognitive problem-solving skills training, or their combination) (Kazdin, 2010). These treatments are rated quite positively and do not differ in level of acceptability. It is possible that the treatments were equally acceptable. Yet, the means for the treatments are close to the upper limit of possible scores on the scale. Thus, it remains possible that acceptability would differ if the ceiling of the scale were not so limited.

There are obvious and subtle problems to be aware of with regard to floor and ceiling effects. The obvious one I have been discussing, namely, is that there may be a numerical limit to the scale that will not allow changes to be found.

This is obvious because when one plots the scores (individual scatter plot) or looks at characteristics of the distribution (mean, variance, skew), the bunching of scores to one extreme is easy to discern.

The more subtle version of ceiling and floor effects is that toward the upper (or lower) ends of a scale (ceiling and floor, respectively), the amount, level, or magnitude of change needed to move one's score at one point on the scale or measure may be much greater than what is needed to change when at a less extreme point on the scale. Let me say this in a way to convey the point, even though it is not quite accurate. When dieting, the first 10 pounds (4.53 kg) is easier to lose than the second 10 pounds. The point is that an increment or decrement of 10 is not equally easy at all points on the measure (in losing weight). For a measure of psychological adjustment or some other construct used in a multiple-treatment design, the first intervention may move people to a mean score of 75 on a scale that goes to 100. The investigator may say there is no ceiling effect because the next treatment still has a lot of room (25 points) to get to the maximum score, so there is no ceiling effect problem. Even so, there may be a ceiling problem. Changes in these last 25 points (from 75–100) may be much more difficult to make than changes from 50–75. There may be a de facto ceiling that the numerical limit does not necessarily reflect. To obtain a score of 90–100, for example, this really requires amazing adjustment that few people would have.

Related, even if it is equally easy to move from one score to another throughout the full range of the possible scores, it may be the case that in fact the extremes of the scale are rarely used by anyone. Here again if one looks at the scores (means and range for the sample), it will look like there is no ceiling or floor problem because there is still room to move further toward each end. Yet, there might be a de facto ceiling or floor effect here. The "real" range of the scale is what many subjects use in fact and the fact that higher or lower scores are possible (but never used) is deceiving.

> How can one tell if these more subtle versions of ceiling effects are likely? One can look to other research that has used the scale. Does other research show that people in fact can or often do score at the extremes of the scale? Is there evidence that the end points provide useful and usable data? If yes, then ceiling (or floor) effects are not likely to be a problem.

Most of the time, it is useful just to worry about the obvious way in which ceiling and floor effects are evident, but the subtle way is not trivial. Support for one's hypotheses require many conditions to coalesce including whether the measure(s) can show group or condition differences when differences really exist.

Although the problem of ceiling and floor effects can occur in any design, it is exacerbated by multiple-treatment designs because different treatments operate toward continued increases (or decreases) in scores. With some manipulations that focus on skills (e.g., reading, musical or athletic skills, something involving practice), the scores may build on each other or accumulate so the main change (from the preassessment) will be evident only for the first condition presented. The second condition, whichever one it is, may not reflect change on the measure very much. Thus, one consideration in using a multiple-treatment design is whether the different measures provide a sufficient range of scores to allow continued increments in performance from a purely measurement standpoint.

From multiple-treatment designs, ceiling and floor effects are readily avoided when behavior change is transient. For example, interventions based upon

administration of drugs or incentives for performance in a decision making game in a laboratory may produce effects only while the condition is in effect. Assessment can be made while these interventions are in effect. After withdrawal of the condition, perhaps with an intervening period so that drug or incentive effects are completely eliminated, the second condition can be implemented. If the effects of a condition are transient and only evident when the condition is in effect, the improvements resulting from one condition will not limit the scores that can be achieved by the second treatment. The study of conditions or interventions with transient effects resolves the problem of ceiling and floor effects in the dependent measures.

Summary and Conclusions: Experimental Research Using Group Designs

Fundamental issues in research are the selection of subjects and their assignment to conditions within the experiment. Random selection was discussed as one possible way of selecting subjects; this is rare in psychological research. Subjects are selected from available samples or those with a particular characteristic (e.g., clinical disorder). Critical sampling issues were discussed, including the heavy reliance on a narrow range and type of subjects. College students have been disproportionately used as subjects in research. Some of that has changed in light of increased evidence that generality of results from such samples is in question for basic psychological processes. Also, access to additional samples with easy access (e.g., from the Web) has expanded the pool. In general, selecting a diverse sample is a default position. That increases the onus on us as investigators to explain more explicitly why a non-diverse sample is a more appropriate or better test of our hypotheses. In clinical research of course, the sample often is determined by patient characteristics of interest. Here too diversity is no less important. In selecting subjects, samples of convenience were also mentioned. These are subjects identified merely because they are available and occasionally because they are serving some other purpose. The overall point is that whatever sample is used, the investigator ought to state explicitly the rationale for using a particular sample and any exclusions that were imposed.

Careful attention must be given to the assignment of subjects to groups. In experimental research, subjects are assigned in an unbiased fashion so that each subject has an equal probability of being assigned to the different conditions. Typically, random assignment is employed. As an adjunctive procedure, subjects may be matched on a given variable at the beginning of the experiment and randomly assigned in blocks to conditions. Matching followed by random assignment is an excellent way to ensure equivalence of groups on a measure that relates to the dependent variable. Matching also was discussed in the context in which subjects are preassigned to some condition by virtue of their status, experience, or some other event not under control of the experimenter. Propensity score matching was highlighted as a method used increasingly to accomplish matching on multiple variables.

Many different designs were discussed including the pretest–posttest control group design, posttest-only control group design, Solomon four-group design (including the Kazdin-on-the-cheap-three-group version), factorial designs, and quasi-experimental designs. In these designs, each subject receives one condition (treatment, control) and groups are compared to evaluate the impact of the intervention or manipulation. The pretest–posttest control group was noted as particularly advantageous because of the strengths that the pretest provides for demarcating initial (pre-experimental manipulation) scores, evaluating change, and increasing power for statistical analyses. Randomized controlled trials are experiments designed to evaluate an intervention and are frequently used in medicine, psychology, education, criminal justice, rehabilitation, and other areas with applied foci. Usually these designs rely on the pretest–posttest control group design.

Multiple-treatment designs were also discussed. In these designs, each subject receives all of the conditions (e.g., more than one treatment or treatment and control conditions). Separate groups of subjects are used so that the different treatments can be counterbalanced. Counterbalancing is designed to ensure that the effects of the treatments can be separated from the order in which they appear. In the simplest multiple-treatment design, referred to as the crossover design, two treatments are given to two groups of subjects but in different order. In more complex versions, three or more interventions or conditions may be delivered and presented either in a randomized or in a prearranged order to randomly comprised groups of subjects. A Latin Square design refers

to ways of arranging multiple treatments within subjects where the number of treatments is equal to the number of groups and where each treatment appears once in each position in the sequence in which treatments are arranged.

There are several considerations in using multiple-treatment designs. Order and sequence effects can emerge and must be controlled by ensuring whenever possible that each treatment is administered at each point in the order of treatments (e.g., first, second, third, etc.). Also, ceiling and floor effects are more likely in multiple-treatment designs, i.e., upper or lower limits on the response measure that will not allow subsequent interventions to reflect further change in performance.

Critical Thinking Questions

1. What is random assignment exactly, and what is it designed to accomplish?
2. What is the difference between a true-experiment and quasi-experiment?
3. What are some of the advantages of using a pretest, as in a pretest–posttest control group design?

Chapter 5 Quiz: Experimental Research Using Group Designs

Chapter 6
Control and Comparison Groups

Learning Objectives

6.1 Identify a control group

6.2 Recall a no-treatment control group

6.3 Recognize the rationale of using the wait-list control group

6.4 Express the phenomenon of sudden gains as applicable to no-contact control groups

6.5 Examine the role of the placebo effect in nonspecific treatment

6.6 Evaluate the ethical considerations in administering treatment as usual

6.7 Report the utility of a yoked control group

6.8 Explain how nonrandom assigned or nonequivalent control group help rule out specific rival hypotheses

6.9 Identify some of the main deliberations while selecting a group

6.10 Assess how intervention research addresses the various research concerns

6.11 Evaluate three additional psychosocial strategies that can be used to develop effective interventions.

In a first introduction to research methods, we are often taught that an experiment requires a control group. Of course, the notion of a control group is mildly misleading because it implies that the addition of a single group to a design may provide a general control for diverse biases, artifacts, and rival hypotheses that might plague the research. In fact, there are all sorts of groups that may be added or included in a design depending on the potential influences other than the manipulation or intervention that may account for the results (threats to internal validity) and the specificity of the statements the investigator wishes to make about what led to change or group differences (threats to construct validity). Indeed, a more in-depth understanding of research design might be pursued by considering the broader concept of comparison groups.

> *Comparison groups refer to any group included in design beyond the primary group or groups of interest.*

Comparison groups permit the investigator to draw various conclusions; the groups differ in the types of conclusions they permit. There might be multiple treatment groups, for example, that differ on one or two ingredients or component because the investigator wishes to dissect treatment or make claims about a particular ingredient. Here the comparison groups all include treatment and in some sense are not "control" groups as this is usually discussed.

The broad issue for any group study is what groups to include in the study, whether variations of some experimental manipulation or an effort to rule out various threats to validity. As one begins to design a study, selection of groups is guided by what one wishes to say at the end of the study. Would you be able to say if the results came out as predicted, or will there be some ambiguity due to internal or construct validity threats?

What do you think?

Whether a study is a well-designed study is determined in large part by whether the investigator is entitled to say what she said in light of the findings. It may be that the findings are fine but the investigator says that they cannot be explained by this or that other interpretation, but the design did not allow for that. Stated in other ways, threats to validity (e.g., internal or construct in particular) may interfere with the conclusions reached by the author. And we then say, the study

was not a poorly designed study necessarily. It may have included all sorts of wonderful practices (e.g., random everything, matching, multiple measures, and more). The "but" is that the conclusions must be connected to the design features. Comparison groups might be needed in the design to allow the author to make the interpretation he or she wishes to advance.

Comparison is the broad and generic category and includes any group that will help reach inferences of interest to the investigator. Comparison groups may include two or more active experimental groups (comparing mood inductions, tasks, priming). Control groups are merely one type of comparison group included in a study. Some control groups (e.g., no treatment, wait list) primarily address the threats to internal validity; other control groups (e.g., nonspecific treatment) address threats to construct validity in the sense that they aid in interpreting the basis for the impact of the intervention. The investigator may wish to make any number of statements about the experimental manipulation and what accounted for the change. Because the range of possible conclusions varies widely with content area and investigator interest, all groups of interest cannot be catalogued. Nevertheless, comparison groups often used in clinical research can be identified and illustrated. This chapter discusses groups that are often used in clinical research, the design issues they are intended to address, and considerations that dictate their use.

In this chapter, we discuss control and comparison groups in the context of intervention studies in part because the range of options is well developed and one can convey how they operate to allow different types of conclusions. The discussion focuses primarily on true-experiments where the investigator is manipulating or controlling the delivery of the intervention. Control groups will emerge again in the context of observational studies because the groups in those studies raise entirely different challenges in addressing threats to validity.

6.1: Control Groups

6.1 Identify a control group

Basic control groups usually are used to address threats to internal validity, such as history, maturation, selection, testing, and others. Control of these threats is accomplished by ensuring that one group in the design shares these influences with the intervention group but does not receive the intervention or experimental condition. If the intervention and control groups are formed by random assignment and assessed at the same point(s) in time, internal validity threats are usually addressed. In clinical research, several control groups are often used. We discuss many different control groups many of which are routinely used in intervention research. Table 6.1 lists the groups we discuss for easy reference and brief summaries.

Table 6.1: Basic and Not-So Basic Control Groups in the Context of Intervention Research Evaluation

Control Group/Strategy	Basic Requirements
No Treatment Control Group	No intervention is provided to this group and assessments (pre/post) are obtained over the same interval that the intervention is provided to the intervention group.
Wait-List Control Group	Identical to the no-treatment control group except after the second (post) assessment, participants now receive the intervention. Sometimes an additional assessment is provided after the wait list group receives the intervention.
No-Contact Control Group	No intervention is provided, but participants are not aware that they are participating in a study and have no contact with staff involved in the project.
Nonspecific Treatment or Attention-Placebo Control Group	An "intervention" is provided to include common factors or nonspecific features of the intervention (e.g., credible rationale, some procedures, participation or attendance) that are not considered to be the critical components that are responsible for intervention effects achieved in the intervention group. This is equivalent conceptually to a psychological "placebo," ergo the alternative name of "attention-placebo" control group.
Treatment as Usual	The standard or routine treatment that is provided in a given setting for the same clinical problem or intervention focus. This group receives whatever is usually done, i.e., as usual care.
Yoked Control Group	Not necessarily a separate group from one of the prior groups in this table. Yoking refers to matching subjects in different groups on some variable (e.g., duration or number of sessions) that might emerge during the course of the study. The investigator wishes to rule out the impact of these probably ancillary differences between intervention and nonintervention groups. Subjects in intervention and nonintervention groups are paired as "partners" so to speak and the emergent variable (e.g., more sessions) that was provided to the intervention subject is assigned to the partner.
Non-Randomly Assigned or Nonequivalent Control Group	A group might be added to the experimental design even though this group could not be randomly assigned to various conditions as were subjects in the other groups. Even so, this group may be selected and used to make implausible specific influences that are not otherwise well controlled (e.g., testing, maturation).

6.2: No-Treatment Control Group

6.2 Recall a no-treatment control group

In evaluating an intervention, a basic question can always be raised, namely, to what extent would persons improve or change without the intervention? *A no-treatment control group in the experimental design receives all of the assessments but no intervention and addresses this question.*

6.2.1: Description and Rationale

> The term "no-treatment group" is commonly used even when interventions are not psychosocial treatments.

This group is so fundamental to intervention research that it was included in the basic descriptions and diagrams of the pretest–posttest control group and posttest-only control group designs. This is the group composed by randomly assigning subjects to conditions (intervention, no-treatment). This latter group is assessed but otherwise receives no intervention. By including a no-treatment group in the design, the effects of history, maturation, and repeated testing as well as other threats to internal validity are directly controlled.

The performance of persons in a no-treatment control group can change significantly over time due to:

- History
- Maturation
- Testing
- Statistical Regression

In some cases, clear historical events are possible explanations. For example, people who are assigned to no-treatment may seek other treatments at another clinic. Even if another type of treatment is not formally sought, clients may improve as a function of more informal means of "treatment," such as talking with relatives, neighbors, members of the clergy, or general practice physicians. Improvements over time may also result from changes in the situations that exacerbated or precipitated the problem (e.g., adjustment after a death of a loved one) and other maturational influences that affect one's mood, outlook, and psychological status. Individuals who come for treatment may be at a particularly severe point in their problem. Hence one would expect that reassessment of the problem at some later point would show improvement. Statistical regression is one explanation for that, but also even without that change over time is "normal." It may not be true that "time heals all wounds," but ordinary processes occurring with the passage of time certainly are strong competitors for many therapeutic techniques. From a methodological standpoint, the important issue is to control for the amount of improvement that occurs as a function of these multiple, even if poorly specified and understood, influences.

> A no-treatment control group assesses the base rate of improvement for clients who did not receive the treatment under investigation.

It is important to use as a no-treatment group clients who have been randomly assigned to this condition. Violation of random assignment erodes the interpretability of any between-group differences after treatment has been completed. For example, some individuals for one reason or another may choose not to participate in the program after pretreatment assessment or withdraw after a small number of treatment sessions. Persons who have withdrawn from treatment would of course not be appropriate to consider as part of or additions to the no-treatment control group. While these clients might be considered to have received no treatment, they are self-selected for that status. Their subsequent performance on any measures might well reflect variables related to their early withdrawal, rather than to the absence of treatment.

6.2.2: Special Considerations

Using a no-treatment control group presents obvious ethical challenges. When clients seek treatment, it is difficult to justify withholding all attempts at intervention. Providing an experimental or exploratory treatment that is reasonable, even if unproven, usually is more ethically defensible than providing no treatment at all. One has to say "usually" because interventions occasionally make things worse, so one cannot assume that a new intervention (therapy, medication, and educational program) will have no effects or positive effects. Even so, no-treatment definitely withholds treatment and that notion is objectionable because it means not even trying to help.

When it comes to withholding treatment in a clinical situation, ivory tower pleas for experimental elegance, control groups, and the importance of scientific research may be unpersuasive to prospective clients. Actually, the ethical issue usually is circumvented by conveying at the outset of a study that one could be assigned to a no-treatment condition and that individuals ought to participate in the study only if this possibility is acceptable. Solicitation of consent to participate in advance of the study conveys the options to the prospective participants and allows them to decide if participation is reasonable. Even this is by no means a great solution. Clients may agree to participate but then end up dropping out if they are not assigned to the treatment group. Indeed, informed consent explicitly specifies that a participant can withdraw at any time. Withdrawing once assigned to a no-treatment group is reasonable or at least quite understandable from

the standpoint of a client. Leaving this issue aside, for clients who are suffering significant impairment or dysfunction and indeed who are in crisis, it is unclear whether assignment to no-treatment would be ethically defensible even if they agreed to participate and remain in the condition.

Aside from ethical issues, there are obvious practical problems in utilizing a no-treatment control group. Difficulties are encountered in explaining to clients who apply why treatment is unavailable or why there is a no-treatment condition. When the study begins, persons who are assigned to the no-treatment condition may seek treatment elsewhere or they may resent not receiving treatment and fail to cooperate with subsequent attempts to administer assessment devices.

If a no-treatment group of clients is successfully formed, it is likely that there will be time constraints on the group. As a general rule, the longer that clients are required to serve as no-treatment controls (interval from pretreatment to the second or posttreatment assessment), the more likely they will drop out of the study. The investigator may wish to know the effects of the intervention over an extended period (e.g., 1, 5, or 10 years of follow-up). However, continuation of a no-treatment condition usually is not feasible over an extended period (e.g., months). Few no-treatment subjects are likely to remain in the study over an extended period; those who do may be a select group whose data are difficult to interpret.

A partial solution to withholding treatment and meeting the requirements of a no-treatment control group is to use a wait-list control group.

6.3: Wait-List Control Group

6.3 Recognize the rationale of using the wait-list control group

Rather than withhold treatment completely, one can merely delay treatment. A *wait-list control group withholds treatment for a period of time after which treatment is then provided.*

6.3.1: Description and Rationale

The period for which treatment is withheld in a wait-list control group usually corresponds to the pre- to posttreatment assessment interval of clients in the treatment condition. Thus, treatment and wait-list cases are assessed at the beginning of the study (before any treatment is given) and at that point when the treatment group has completed treatment. The wait-list group will not have received the intervention during this period but will have completed all of the pre and "post" assessments (it is really a second pretreatment assessment for them because they have not received treatment). As soon as the second assessment battery is administered, these subjects can begin treatment.

> When clients originally apply for treatment, they can be asked whether they would participate even if treatment were delayed. Only those subjects who agree would be included in the study. These clients would be assigned randomly to either treatment or wait-list control conditions.

The control clients are promised treatment within a specified time period and in fact are called back and scheduled for treatment. Although it is tempting to assign those clients who indicate they could wait for treatment to the control group and those who could not wait to the treatment group, circumventing random assignment in this way is methodologically disastrous. Treatment effects or the absence of such effects could be the result of subject selection in combination with history, maturation, regression, and other threats to internal validity.

The three rudimentary features that characterize a wait-list control group are:

1. If a pretest is used, there must be no treatment between the first and second assessment periods for the wait-list control group. During this period, the group is functionally equivalent to a no-treatment control group.
2. The time period from first to second assessment of the wait-list control group must correspond to the time period of pre- and posttreatment assessment of the treatment group. This may be easily controlled if treatment consists of a particular interval (e.g., 2 months) and the pre-to-posttreatment assessment period is constant across treated subjects. Then, wait-list control subjects can return for reassessment after that fixed interval has elapsed. If treatment duration varies, a wait-list subject might be reassessed at the same interval of the treatment subject to which he or she has been matched (elaborated further later in the discussion of yoking). For example, a wait-list control subject can be scheduled for reassessment at the same time when a treated subject returns for posttreatment assessment. The wait-list control and experimental subjects are grouped in this way on the basis of having taken the pre- and posttreatment assessment devices over the same time interval (e.g., within 1 week), or perhaps even on the same days. It is important to keep the time interval constant to control for history and maturation over the course of the assessment interval.
3. Waiting-list control clients complete pretest or posttest assessments and then receive treatment. An important practical question is how to have the wait-list subjects return for reassessment immediately prior to providing them with treatment. Actually, this is not particularly difficult. Clients usually are required to complete the assessment again before receiving treatment.

Essentially, reassessment is connected with the promise of scheduling treatment and serves as an immediate antecedent to the long-awaited intervention.

6.3.2: Special Considerations

Use of a wait-list control group has much to recommend it. From a practical standpoint, it usually is not as difficult to obtain wait-list control subjects as it is to obtain no-treatment subjects. The difficulty partially depends upon how long the controls are required to wait for treatment, the severity of the problem, their perceived need for treatment, and the availability of other resources.

From the standpoint of experimental design, there is a decided advantage in the use of a wait-list control group. This group allows careful evaluation of treatment effects at different points in the design. Because treatment eventually is provided to the wait-list control subjects, its effects can be evaluated empirically.

Essentially, a wait-list control study using a pretest can be diagrammed as follows:

$$\begin{array}{llll} R & A_1 & X & A_2 \\ R & A_1A_2 & X & A_3 \end{array}$$

The effect of treatment (X) is replicated in the design. Not only can the treatment be assessed by a between-group comparison (comparison of A_2 for each group using A_1 as a covariate or part of repeated measures analysis) but also by within-group comparisons as well (comparison of change from A_1 to A_2 separately for each group and the change from A_2 to A_3 for the wait-list group by within-group *t* tests). Of course, to accomplish this, wait-list control group subjects must be reassessed (A_3) after they finally receive treatment.

The wait-list control group does not completely ameliorate the ethical problems of withholding treatment but may help a little. Now the issue is not withholding treatment from some of the clients. Rather, all clients receive treatment and differ only according to when they receive it. Ethical problems arise if clients request or require immediate treatment and delaying treatment may have serious consequences. Obviously, a wait-list control group is not ethically defensible with patients at risk for injury or death (e.g., self-injury, suicide) but also if they are suffering or impaired by their own account or by other measures (e.g., family report, initial screening assessment).

Apart from such situations and as an alternative to the no-treatment control group, a wait-list group offers a distinct advantage because clients eventually receive treatment. That also gives the possibility of demonstrating replication of treatment effects as the wait-list group can eventually provide outcome data once they finally complete treatment.

There is a limitation of the wait-list control condition. Because subjects in the wait-list group receive treatment soon after they have served their role as a temporary no-treatment group, the long-term impact of such processes as history, maturation, and repeated testing cannot be evaluated. Even if wait-list subjects did not change very much in the time interval in which they waited for treatment, they may have improved or deteriorated greatly by the time of follow-up assessment even without treatment. One can follow the treatment group to see how they are doing 1 or 2 years later. Yet, the wait-list control group is no longer available for comparison; by this time, this group will be another treatment group. It is important to be aware of this disadvantage. If one is planning follow-up in an intervention study, this is certainly relevant.

6.4: No-Contact Control Group

6.4 Express the phenomenon of sudden gains as applicable to no-contact control groups

The effects of participating in a study, even if only in the capacity of a no-treatment or wait-list control subject, may have impact on the subjects because participation can be reactive. In the context of treatment research, participating in a control group may exert some therapeutic change. Indeed, in the early history of psychotherapy research, we learned that clients who only receive the initial assessment battery on separate occasions before any treatment begins show marked improvements (Frank, Nash, Stone, & Imber, 1963). Among the interpretations, one is that just the process of entering a treatment study mobilizes hope and expectations for improvement but certainly statistical regression is as if not more plausible and parsimonious.

Fast forward to contemporary research and we know about a phenomenon referred to as *sudden gains* (Aderka, Nickerson, Bøe, & Hofmann, 2012). This refers to changes made very early in treatment and sometimes even before the putative critical ingredients of the intervention have been provided. For example, specific cognitive processes or special activities may be viewed by the investigator as the essential components of treatment and on which change depends. Many clients may change suddenly and even before these components are provided. The sudden gains can be as enduring as those achieved with the full treatment regimen. In short, something might be going on by participating in treatment or a treatment study.

Occasionally it is possible to evaluate the impact of participation in a study by using as a control group individuals who have no contact with the project. These individuals constitute a *no-contact control group*.

6.4.1: Description and Rationale

The requirements for a no-contact group are difficult to meet because the subjects do not receive treatment and do not realize that they are serving in this capacity. To obtain such a group of subjects, pretest information usually is needed for a large pool of subjects who are part of a larger assessment project. Some of these subjects, determined randomly, are selected for the no-contact group. The initial assessments are administered under some other guise (e.g., part of routine class activities in an undergraduate course). Also, obtaining subsequent test information must be conveyed as part of a routine activity, so it is not associated with a treatment project. Testing on measures relevant to a psychological study might be part of a routine class activity or some other purpose (e.g., administering tests to all introductory psychology students as part of a subject pool that might be used for research or testing of all college athletes that is routine or standard). In each case, assessment is disconnected to a study or at least a specific study.

A no-contact control has come to have different meanings over time. In the initial use and still relevant methodologically, it meant that individuals are not aware that they are involved in a study. The classic example focused on treating speech anxiety among college students (Paul, 1966). Several students who qualified for treatment were used as no-contact control subjects. Measures were administered under the guise of requirements for and a follow-up to ordinary college speech classes. To clarify, two control groups were included in the study:

1. There was a no-treatment group. These subjects were aware they were in the study, received several assessment devices as part of the study, telephone contact and interviews, and other procedures related to the treatment project.
2. There was a no-contact group. These subjects completed assessments before and after "treatment" of other subjects, but these assessments were part of participation in the speech class. There was no phone contact with these individuals. Thus, data were available without revealing use of the information as part of a treatment study.

At the end, comparisons could be made assessing the effect of receiving contact with the program (no treatment subjects) versus no contact. Among subjects who did not receive treatment, those who had no explicit contact with the study (no-contact controls) performed less well on various measures of anxiety and personality at the end of the study and at follow-up than those subjects who did (no-treatment controls who knew they were part of the study). Thus, serving as a no-treatment subject explicitly connected with the study was associated with some improvements that did not occur for no-contact subjects. This is instructive because it conveys that knowledge of participation (reactivity) can lead to change. The term "no-contact group" is used in another sense than the control procedure I have described.

> In treatment research, an increased focus is the use of self-help interventions. With these interventions, individual clients take control and implement treatment for themselves.
>
> There are many variations that reflect a continuum of external support and contact, including complete independence (no contact); group support; and minimal to full-time aid from volunteer, semiprofessional, or professional help (Harwood & L'Abate, 2010). Self-help interventions use various media (i.e., apps, Web-based interventions, videos) to provide treatment. Although many mental health problems have been studied, anxiety and depression have received the greatest attention. In some of these studies, the intervention group (rather than a control group) involves no contact.
>
> In this context, no contact refers to the intervention group where participants receive no contact with a therapist or mental health professional. This is different from the prior use of the term I noted where no contact meant not knowing one was even participating in a study.

For example, in one study, clients with panic disorder were assessed and then assigned to one of two conditions: bibliotherapy or wait-list control group (Nordin, Carlbring, Cuijpers, & Andersson, 2010). Bibliotherapy consisted of providing a self-help course of cognitive behavior therapy strategies for the treatment of panic. All subjects knew they were in a study. In this case, no contact meant the subjects carried out treatment without therapist contact or assistance for the 10 weeks of treatment. The group improved significantly better on outcome measures immediately after treatment and 3 months later when compared with the wait-list control. In this study, no contact is not a control group but rather the treatment group. In that sense, this is not what is meant by no-contact controls in the present chapter. Yet, it is important to mention the diverse uses of "no contact" to convey the term does not always mean a control condition. One of the challenges in provided mental health services on a large scale is that it is difficult to get treatment to people in need for a variety of reasons (e.g., too fewer therapists in places such as rural areas, stigma of going to therapy). Use of treatments with minimal or no therapist contact is one of many strategies to try to improve the reach of therapy (Kazdin & Blase, 2011).

6.4.2: Special Considerations

Use of a no-contact control group where participants are not aware that they are participating in a study is rarely a viable option for research. Administering measures under a guise other than participating in a study is likely to violate both the letter and spirit of current informed consent

requirements for participants in research. Studies in institutional settings such as schools, clinics, hospitals, and prisons and studies with large populations engaging in standardized testing such as the military or all entering high school or college students might permit delineation of a no-contact control group. Assessment devices could be administered routinely on separate occasions and be used to provide data for comparisons with subsamples that serve in the study. And special matching techniques (e.g., propensity score matching) could be used to get a matched no-contact group to serve as a basis of comparison with an intervention group. Even so, use of data as part of research almost always requires informing subjects and obtaining consent. Exceptions can be made when identity is completely obscured (e.g., anonymous records review) but that is not always the case. As a general rule, subjects would need to be informed that they are serving in a study, perhaps even more so when an intervention is involved and the potential implication that someone needed treatment.

> The main issue is not in whether a no-contact group could be formed but rather the requirements of the research question. In most studies, the investigator is not likely to be concerned with separating the effects of contact with the treatment or research project from no-contact; a no-treatment or wait-list control group is likely to serve as the appropriate measure of improvement against which the effects of treatment can be evaluated.

On the other hand, it might be important for conceptual reasons to evaluate whether serving in a project, even as a no-treatment subject, influences a particular set of measures or clinical problem. A fundamental issue in developing effective interventions (e.g., psychological treatment, educational interventions) is that the interventions (e.g., procedures, techniques) are the critical ingredient. In that context, it is useful to know how much change occurs anyway if individuals believe they are participating in a study (e.g., no-treatment or wait-list group) versus not believing they are participating in a study (e.g., no contact). In addition, understanding how changes come about without interventions would be informative and potentially useful to harness.

6.5: Nonspecific Treatment or Attention-Placebo Control Group

6.5 Examine the role of the placebo effect in nonspecific treatment

The next type of control group we will examine is the nonspecific treatment or attention-placebo control group.

6.5.1: Description and Rationale

No-treatment and wait-list control groups are employed primarily to address threats to internal validity (e.g., history, maturation, repeated testing). In the context of treatment research, a nonspecific-treatment control group not only addresses these threats but also focuses on threats to construct validity. In any treatment, there are many seeming accouterments that may contribute to or be responsible for therapeutic change. Such factors as attending treatment sessions, having personal contact with a therapist, hearing a logical rationale that describes the supposed origins of one's problem, and undergoing a procedure directed toward ameliorating the problem may exert influence on client performance and generate their own therapeutic effects. These factors are referred to as *common or nonspecific factors* of psychotherapy because they are ingredients in most treatments. Moreover, when we consider specific therapy techniques (e.g., cognitive behavioral treatment, multisystemic therapy), we usually do not know the mechanisms of action or processes through which they achieve their effects. It might be specific facets of the procedures (e.g., activities and exercises directed toward change) or due to these common factors, or some combination.

Common factors may be critical to psychotherapy because of the processes they mobilize within the individuals and the changes those processes produce. When clients participate in treatment, they are likely to believe in the procedures and have faith that some therapeutic change will result (Lambert & Ogles, 2013). We have learned from many years of medical research that the belief in treatment is important. Placebos, inert substances (e.g., sugar tablets), given under the guise of treatment can alter a variety of disorders ranging in severity from the common cold to cancer (e.g., Benedetti, 2009; Finniss, Kaptchuk, Miller, & Benedetti, 2010). Even more than that we know that some placebos (e.g., active placebos that have side effects similar to medication) and some ways of administering placebos (e.g., larger pills vs. smaller ones; injections rather than pills) even increase or strengthen placebo effects.

> Placebo effects, by definition, result from factors other than active ingredients in the substance itself. Hence the belief of the patient in treatment and perhaps the belief in the physician who administers treatment and similar factors appear to be responsible for change.

Effects analogous to placebo reactions influence individuals who come to psychotherapy. Indeed, the history of psychological treatments can be traced by drawing attention to procedures and therapists (e.g., Franz Anton Mesmer [1734–1815] and Emile Coué [1857–1926]) whose effects we recognize to have been largely due to suggestion. With the perspective of time, these procedures and their inventors have been dismissed, but the effectiveness

of their procedures is not in question as much as the reasons they used to explain the effects. What all this means for the present discussion is that in an empirical investigation of psychotherapy, a simple comparison of treatment and no-treatment control groups does not establish what facet of "the intervention" led to change, i.e., construct validity. To identify if the specific intervention or the unique properties of a treatment are important in producing change in the clients, a nonspecific-treatment group can be included in the design.

6.5.2: More Information on Description and Rationale

As you may recall, earlier in the chapter I noted that the comparison and control groups one selects is based on what one would like to say at the end of the study. Here is a good example. An investigator may want to show the effects of mindfulness treatment for social anxiety and includes a treatment and a no-treatment control group. So far so good. At the end of the study (in the Discussion section), the investigator can talk about treatment being more effective than no treatment, on the assumption that is the pattern of results. However, the investigator may slip into something more comfortable such as talking about why mindfulness as a technique worked or how mindfulness overcomes core elements of the disorder. These latter points are not what the design allows the investigator to say. Construct validity and specifically the possibility of common factors explaining everything were not addressed in the design. Is the design flawed? Not at all. This is a fine or poorly designed study based on what the investigator wants to say.

To address a critical construct validity issue in treatment studies, a nonspecific-treatment control is one option. That group may include procedures in which clients meet with a therapist, hear a rationale that explains how their problem may have developed, and discuss something about their lives in sessions that are similar in number and duration to those in the treatment group. From the standpoint of the investigation, these subjects are considered to be receiving a psychological placebo, as it were. This is some procedure that might be credible to the clients and appears to be effective but is not based on theoretical or empirical findings about therapeutic change.

In developing a nonspecific control condition, the goal is to provide some form of pseudo intervention that involves clients in an experience of some sort. The goal is to control for factors common to coming to treatment but without the putatively critical ingredient. Special care is needed to decide in advance what the investigators wish to control. For example, if one wants to say that the intervention was responsible for change and not just the common factors, then the control group is the one we have been discussing in which attending sessions, meeting with a therapist, and the like are included. However, if one wants to say more, such as the treatment led to change because of how cognitions were addressed or whether mindfulness makes a difference, this is not handled by a nonspecific treatment control group. In this case, one might use a nonspecific treatment control group but also add a comparison group that provides the "real" treatment minus some putatively procedures or ingredients (e.g., cognitive procedures or mindfulness).

A **nonspecific-treatment control group** is designed to control for common factors that are associated with participation in treatment. If a treatment group is shown to be more effective than a nonspecific-treatment control group, this does not necessarily mean that the processes proposed by the investigator to characterize the treatment group (e.g., changes in cognitions, resolving conflict) are necessarily supported. Nonspecific-treatment control groups rule out or make implausible some common factors as an explanation of the results, but they do not necessarily point to the construct in the treatment group that is responsible for change. If the investigator wishes to argue for the basis for change in the treatment group, some evaluation of the processes considered to be central to change (e.g., cognitions, alliance) ought to be assessed directly and tested in relation to the amount of therapeutic change.

6.5.3: Special Considerations

There are several issues that emerge in developing a nonspecific treatment control condition. To begin, the conceptual problems are not minor.

What is an inert intervention that could serve as a control?

A placebo in medicine is known in advance, because of its pharmacological properties (e.g., salt or sugar in a tablet), to be inert (not to produce effects through its chemical properties in relation to the clinical problem). In psychological treatment, one usually does not know in advance that the properties of the nonspecific-treatment group are inert. Merely chatting with a therapist or engaging in some activities vaguely related to one's problem might be cast in theoretical language to make them seem plausible as genuine treatments. This is why in the therapy research business, one investigator's treatment group is another investigator's control group. It is difficult to devise an intervention that is at once credible to the clients and yet one that could not also be construed by someone as a theoretically plausible treatment as well.

Another issue that emerges pertains to the credibility of the procedure. One ingredient in therapy is the client's beliefs or expectancies that treatment will work. Presumably a plausible nonspecific-treatment control group would have

this ingredient as well so that client expectations for improvement could not explain outcome differences between treatment and control conditions. However, devising a credible control condition requires a rationale about why this "treatment" is likely to be effective and why procedures in or outside of the treatment sessions look like credible means toward therapeutic change. It may be the case that the control condition is not as credible and does not generate expectancies for change as well as the veridical treatment group to which it is compared. Indeed, highly credible control conditions are often just as effective (when compared to no treatment) as treatment conditions (see Lambert & Ogles, 2013). In fact, the more the attention control condition generates expectancies for improvement that approach or equal those of the intervention group, the less likely there will be differences between the two conditions (see Baskin, Tierney, Minami, & Wampold, 2003; Boot, Simons, Stothart, & Stutts, 2013; Grissom, 1996).

From a methodological perspective, the similarity of credible control conditions and treatments has implications for conducting experiments (larger sample sizes are needed to detect small group differences) and for their interpretation (isolating the construct that accounts for change). Also, measures can be used at the beginning of treatment to assess treatment credibility and client expectancies for improvement. For example, clients can be asked questions about how credible treatment is, how logical it appears, and how likely treatment is to be successful. Responses to such items can be evaluated in relation treatment outcome (e.g., is therapeutic change a function of initial credibility of the treatment?). The data can be brought to bear on the likelihood that expectancies or differential expectancies play a role in treatment outcome.

6.5.4: Ethical Issues

Ethical issues also emerge in providing nonspecific-treatment conditions, beginning with the problem of providing a treatment that is not well based on theory or empirical findings. In addition, if clients are in need of clinical care, this type of group may not be defensible. The ethical issues have become even more salient in light of current developments in the ethics of medical research. Although we take up ethical issues in greater depth later, a key point is pertinent now.

Research guidelines include many professional codes, but one is worth mentioning in the context of the present discussion. The Declaration of Helsinki is a major international code of ethics for biomedical research involving human subjects devised by the World Medical Association (see Carpenter, Appelbaum, & Levine, 2003). The declaration and its guidelines for research were prompted by gruesome medical experiments of the Nazi era and designed to protect subjects. The declaration began in 1964 and is periodically revised. In the guidelines, placebo control groups are to be used only under very special circumstances and these must be judged to be compelling. The comparison intervention ought to be the best current treatment available as the default control group condition.

We have been discussing psychotherapy, but much of the controversy about the use of placebo control conditions grew out of medical research on such serious conditions as HIV in which efforts were made in Africa and Asia to evaluate new medications (e.g., to prevent pregnant HIV positive mothers from passing HIV to their newborns, or to prevent sex workers from contracting HIV in the first place). The use of placebos in such trials has been controversial to say the least. Citizens have lobbied against use of placebos when there is a reasonable basis for providing the drug or another treatment.

> Whether or not the research focus is life threatening, patients ought not to be subjected to placebo control conditions if any reasonable alternative conditions could be provided. There is no justification for unnecessary suffering. Arguments on the other side have focused on the need to establish the effects of treatment to ensure the greatest benefit.

The issue may not be resolvable definitively because of the different stages of treatment development for a variety of disorders and whether placebo effect is important to control in any given instance. Internationally, there is no standard practice. The Helsinki guidelines are not binding but voluntary, and some countries and agencies within a country have their own standards.

For psychological research, it is reasonable to ask, ought one to use a nonspecific-treatment control condition, and if so under what circumstances?

From the standpoint of the scientific underpinnings of therapy, the control for common factors is important. Indeed, claims are made that the effectiveness of a specific therapy (over and above client expectancies or common factors) can be argued. Some areas of treatment (e.g., cognitive therapy for depression, dialectical behavior therapy for bipolar disorder) have extensive literature on their effects, and comparisons with other treatments or strong control conditions suggest they have benefits well beyond expectations. Yet, the magnitude of those benefits is not so clear.

Use of a nonspecific-treatment control condition can have deleterious effects on the clients, apart from the absence of immediate benefit for the clinical problem leading them to treatment. Assume for a moment that one is able to devise a nonspecific-treatment control group and provide this to clients. Perhaps the client is not likely to get better, although the rate of improvement will vary as a function of client problem and quality/credibility of the attention placebo group. Participation in the nonspecific-treatment control condition might influence beliefs about therapy in general and have impact on client's subsequent use of treatment.

The client who receives a "fake treatment" might be turned away from therapy in the future when a veridical treatment might help with the stresses and strains of life. The nonspecific-treatment group may not be very credible or does not help the client, and hence leads the client away from a potentially useful resource. Conceivably ordinary therapy might teach a given client such lessons; using a control condition without a veridical treatment merely increases the likelihood of such an effect.

Research to date tends to support the view that psychotherapy is more effective than nonspecific-treatment control conditions and that nonspecific-treatment control conditions are more effective than no treatment (Lambert & Ogles, 2013). At the same time, this has been a difficult area of research because of the obstacles of designing and implementing attention placebo conditions that generate as much expectancies for change as the treatment conditions to which they are compared.

> In developing or evaluating a new treatment, it is critical to show that treatment effects surpass those achieved with the common factors that arise from merely participating in treatment. This can be accomplished by using a nonspecific-treatment control group or another treatment that has already been shown to be effective.

Treatment as usual is a viable comparison group too and is discussed next.

6.6: Treatment as Usual

6.6 Evaluate the ethical considerations in administering treatment as usual

In clinical research, assigning individuals to no-treatment, wait-list, and nonspecific-treatment control conditions may not be ethically defensible or feasible in light of presenting problems of the clients and the context in which treatment is provided. In such circumstances, the investigator may still wish to test whether a new treatment is effective. An alternative that has gained prominence in the past decade is comparing the treatment of interest with the routine or standard treatment that is usually provided at a clinic (Freedland, Mohr, Davidson, & Schwartz, 2011).

> That routine treatment is referred to as *treatment as usual* or *TAU*. At first blush, the term is very clear—the control or comparison will be giving people what they usually would get. Yet, the term is deceptive and arguably just plain odd.

6.6.1: Description and Rationale

Treatment as usual for a given problem (e.g., major depression, bulimia) at one clinic is not at all the same at another clinic and indeed within a given clinic, there is no standard or consistent use of TAU among different therapists. It gets worse—many individuals in practice are wont to say no two patients are alike and that their treatments are tailored to each patient. That means that TAU as administered by a given clinician at a given clinic for two patients with the same or similar diagnoses might be different treatments. TAU is like soup de jour in a restaurant—it may be great but it keeps changing—in this case changing within a clinician, across clinicians, and across clinical settings. In short, we have no clear idea of what TAU actually is (see Kazdin, 2013b). Understandably, methodologists would prefer to call this *treatment as unusual* but that has not gained popularity beyond the dinner table at my family meals.

Yet, TAU is used often in current work and has benefits. As an example, TAU was used as a control group for a study of postnatal depression in women (Mulcahy, Reay, Wilkinson, & Owen, 2010). Depression after giving birth affects approximately 13% of mothers and if not treated can progress to chronic depression. Apart from the suffering of the mothers, depression has deleterious effects on children and family relations more generally. In this study, mothers with postnatal depression were assigned to receive group interpersonal psychotherapy or TAU. Interpersonal psychotherapy focused on dealing with social isolation and feelings of loneliness, receiving sources of social support, and help directly in supporting interpersonal relationships and experiment with new behaviors directed toward these ends. TAU consisted of a variety of community treatments that were routinely available to women, including individual therapy, group therapy, medication, natural remedies, and so on. Women assigned to this condition were given written and verbal information about the local services available. Assessment revealed that TAU participants accessed a range of services.

> **The results**: Both treatment and TAU groups led to significant improvements in depression.

However, the magnitude of changes in the interpersonal psychotherapy treatment was significantly larger, improvements continued after treatment, and greater change was also evident on measures of marital functioning and mother perception of the mother–infant bond. Thus with a strong control group, the benefits of the group treatment were evident. The study is unusual in allowing diverse treatments to be accessed as TAU. This gave participants the benefit of choice in selecting their care.

At least four advantages accrue to the use of TAU as a comparison condition:

1. Demands for service and ethical issues associated with many other control conditions are met. All persons in the study can receive an active treatment and what they might normally receive anyway, i.e., treatment as usual. No one receives a fake condition or procedure that is intended not to work (e.g., a nonspecific-treatment control condition).

2. Because everyone receives a veridical treatment, attrition is likely to be less than if various control conditions were used (e.g., no treatment, wait list). Attrition is implicated in all types of validity (internal, external, construct, and data evaluation) and hence that benefit is not minor.
3. As usual care provided at a facility is likely to control for many of the common or nonspecific factors of therapy (e.g., contact with a therapist, participation in sessions). Thus, receipt of an or any intervention is not a viable rival interpretation of the results in most studies, although "new and improved" therapies when compared to treatments as usual tend to have more enthusiasm, investigator hype, therapist expectations, and novelty effects than business (treatment) as usual.
4. Clinicians who might serve as therapists in the study as well as clinicians who might be consumers of the research results are likely to be much more satisfied with the study that uses standard treatment as a comparison condition. The question is one that is clinically relevant (is the new treatment really better?) and the study more closely resembles clinical work by including a treatment that is routinely used.

6.6.2: Special Considerations

TAUs raise their own dilemmas:

- It is difficult to know what these treatments entail at a clinic, hospital, or school, no matter what the descriptions and brochures actually say.
- The investigator ought to monitor and assess carefully what is done as part of routine treatment.
- It is better from the standpoint of the design for the investigator to oversee, monitor, and evaluate even the TAU so that one can report what was actually done.
- Stated less diplomatically, TAU at many clinics may be administered sloppily, inconsistently, and with great therapist flexibility, personal style, and taste.

(All of this is understandable due to the varied training experiences of the clinicians and their attempts to individualize treatment to the clients, which we do not quite know how to do at this point.) In addition, ethical dilemmas often arise after a study is completed and treatment is shown to be better than as usual care. In such studies (e.g., as cited above with group interpersonal therapy for postnatal depression), routine care may quickly become ethically less defensible because it is shown to be inferior to a new treatment. Usually after such a study, TAU still continues as if the study were not done. This is no fault of the investigator but due to the huge challenges and costs of disseminating findings to change routine care.

There are some scientific dilemmas about TAU. Treatments as usual sometimes work and sometimes are less effective but not by much when compared to evidence-based treatments (Cahill, Barkham, & Stiles, 2010; Freedland et al., 2011; Weisz et al., 2013). And we do not always know that the superiority of evidence-based treatment for a particular problem makes a difference in the lives of the clients when compared to a treatment as usual that has been shown to work. Clearly, there are some scientific issues to work out here to establish where and when evidence-based treatments make a major difference.

Another issue is that by and large treatments as usual are not replicable. That is, they are not documented procedurally (e.g., manuals, guidelines) and they vary in many ways as I have noted. No real conclusions can be reached about them with any generality because the procedure from one clinic to the next may differ greatly. So while TAU solves some control group dilemmas and the ethical issues those control groups raise, we are pretty much at a loss in identifying what TAU is beyond the confines of a particular setting (Kazdin, 2013). That means that the effects of some intervention against a TAU in one setting have no necessary bearing on the effects of that same treatment in comparison to a TAU in another setting down the street.

> The ambiguities of a TAU condition can be rectified by documenting what in fact was done or observing treatment as usual and developing guidelines and procedures from that and making sure that they are followed during a study.

Ironically, carefully structuring, monitoring, and overseeing treatment as usual remove it from the realm of "as usual" where very little of that is going on. With all that said, TAU is a preferred control condition in many ways. The group is likely to circumvent many of the ethical and practical issues of nonspecific-treatment control conditions and to enhance the conclusions that can be drawn about treatment.

6.7: Yoked Control Group

6.7 Report the utility of a yoked control group

Differences in procedures or events to which the subjects are exposed may arise during the investigation as a function of executing the study or implementing a particular intervention. The problem with differences that can emerge is that they are not random but may vary systematically between groups. Essentially, a confound—some variable associated with the intervention—could emerge that might explain the differences between groups. That is, it was not the intervention but this emergent difference between groups that could explain the results. Such differences would need to be anticipated and controlled.

> One procedure to rule out or assess factors that may arise as a function of implementing a particular intervention is called the *yoked control group*.

6.7.1: Description and Rationale

> *The purpose of the yoked control group is to ensure that groups are equal with respect to potentially important but conceptually and procedurally irrelevant factors that might account for group differences* (Church, 1964).

Yoking may require a special control group of its own. More likely in clinical research, yoking can be incorporated into another control group. In this case, yoking refers to equalizing the groups on a particular variable that might systematically vary across conditions. In this sense, yoking is more of a procedure to match intervention and nonintervention subjects and often that can be done with the groups in the design (e.g., intervention vs. nonspecific treatment control group).

Consider a hypothetical study designed to evaluate a specific therapy technique for the treatment of acrophobia (fear of heights). Two groups are used including:

1. The "new and improved" treatment
2. A nonspecific-treatment control group that meets with a therapist but engages in a task not likely to be therapeutic (e.g., discussing the development of fears among people they know)

Suppose that clients in the treatment group are allowed to attend as many sessions as needed to master a set of tasks designated as therapeutic. For example, clients might have to complete a standard set of anxiety-provoking tasks in therapy to help them overcome anxiety. The number of sessions that clients attend treatment could vary markedly given individual differences in the rate of completing the tasks. A nonspecific-treatment control group might receive a bogus treatment in which group members merely come in and discuss fears of their friends and relatives. One might raise the following question.

How many sessions should the control group subjects receive?

It would be important to design the study so that any differences between treatment and control groups at the end of the study cannot be due to the different number of sessions that the groups received or the different elapsed time between first (pre) and second (post) assessments for the groups. The control subjects should not simply be given a fixed number of sessions since that would not guarantee equality of sessions across groups.

A solution is to yoke (match) subjects across groups by pairing subjects. The pairs might be formed arbitrarily unless matching was used to assign subjects to groups. A subject in the experimental group would receive a certain number of therapy sessions on the basis of his or her progress. Whatever that number is would be the number of sessions given (assigned) to the control subject to whom the subject was yoked. That is, the number of sessions for each control subject would be determined by the subject to whom he or she was paired or matched in the experimental group. Obviously, the yoking procedure requires running the experimental subject first so that the number of sessions or other variable on which yoking was done is known in advance and can be administered to the control subject with which the treatment subject is paired. The behavior of the experimental subject determines what happens to the control subject. At the end of treatment, the number of treatment sessions will be identical across groups. Hence any group differences could not be attributable to the number of sessions to which clients were exposed. Yoking would have ensured that the number of sessions did not vary. The yoking would hold constant the number of sessions between the treatment and nonspecific control groups. The yoking might be extended to address the other group in the design, namely, the no-treatment control group.

6.7.2: More Information on Description and Rationale

If a third, no-treatment group were in the design, yoking might be used with that group as well. If pre- and post-treatment assessments are provided, how long should the interval between these assessments be for the group that does not receive any treatment?

What do you think should be the duration between assessments?

Subjects in the no-treatment group could also be yoked to persons in the treatment group in terms of the number of weeks between pre- and post-treatment assessment. Thus, at the end of the study, yoking would yield the following result. Both treatment and nonspecific control groups would have received the same number of sessions and the time elapsed in weeks or days between pre- and post-treatment assessment would be the same for the all treatment and control conditions. The means and standard deviations would not differ for the number of sessions (for the two treatment groups) or the number of days or weeks between pre- and post-treatment among groups. As evident from this example, the yoked control procedure may not necessarily constitute a "new" control group. Yoking often can be added to such a group as a nonspecific-treatment control group.

The importance of yoking can be illustrated in a study designed to improve balance in people with Parkinson's disease (PD; Chiviacowsky, Wulf, Lewthwaite, & Campos, 2012). Individuals with PD are at high risk for falling; many of the falls lead to injury requiring health care services and many of the injuries are fractures requiring surgery. Increasing balance was the goal of the study but with the

primary focus on evaluating a self-control procedure to develop balance. Individuals with PD were assigned to one of two groups:

- The first group required individuals to stand on a platform (stabilometer) while trying to keep their balance (keeping the platform as horizontal as possible during trials). In this self-control group, participants could use a balance pole if they wished to help themselves. They had a choice to use or not use the pole, and that choice was the self-control part of the manipulation.
- The second group also stood on the platform for the same trials. The investigators wanted to test for self-control—selecting the pole for assistance as needed in the first group. Yet, it may not be self-control at all that separated groups, but how many times individuals in each group used the pole as an aid. Perhaps using the pole more as an aid would help with training, leaving aside any choice or self-control in using that pole.

Consequently, participants were yoked in pairs. The number of times a pole was used by a participant in the self-control group was yoked to the partner. That is, the partner in the second group was handed the pole (no self-control, no choice) the same number of times as used by the self-control group partner to whom he was yoked. The results: both groups had an equal number of time and training trials on the platform, and groups were no different in the number of times they used the pole. Yet, the self-control group performed better on the test of balance a day later, after the training had ended. In this study, yoking on use of the pole removed that from explaining why subjects in one group performed better than subjects in another group. Variation in use of the pole would have been a threat to construct validity (i.e., what about the intervention made a difference?). Yoking equalized the opportunities to use the pole.

> Yoking is a way of matching during an experiment if any facet of the intervention or experimental manipulation can vary between groups.

For example, if feedback or reinforcement is provided to participants in one group based on how they perform, this can be controlled by yoking subjects to another group that receives the controlled (same) amount (feedback, reinforcing consequences) of the yoked partner (e.g., Ali et al., 2012; Bickel, Yi, Landes, Hill, & Baxter, 2011). By yoking, the investigator controls those variables that potentially can confound the results.

6.7.3: Special Considerations

Conceivably, an experimental and a control group can be yoked on all sorts of variables that may differ between groups. Whether yoking is used as a control technique needs to be determined by considering whether the variables that may differ across these groups might plausibly account for the results. For example, in a given therapy study it might make sense to yoke on the number of treatment sessions because the amount of contact with a therapist and treatment may contribute to the differences between a treatment and a nonspecific-treatment control group, particularly if therapy subjects receive many more sessions. Stated differently, it may be plausible that the number of sessions, rather than the content of the sessions, is viewed as a threat to construct validity.

> "The intervention" confounds content and amount of treatment and hence raises ambiguities about why the intervention was more effective.

On the other hand, it may be unimportant to yoke subjects in such a way that the time of the day when therapy sessions are held or the attire of the therapists is perfectly matched across groups. The variables that serve as the basis of yoking often are based on considerations of construct validity.

> The usual question for selecting control groups applies, namely, what would the investigator want to say about the treatment effects at the end of the study? It is likely that she would not want the interpretation to be clouded by an emergent variable that systematically differentiated treatment and control groups but was ancillary to the hypothesis. Interpretation of the effects is of course an issue of construct validity. Construct validity is important, and hence control of possible confounds is nothing to yoke (joke) about.

6.8: Nonrandomly Assigned or Nonequivalent Control Group

6.8 Explain how nonrandom assigned or nonequivalent control group help rule out specific rival hypotheses

Many groups might be added to an experiment that utilizes subjects who were not part of the original subject pool and not randomly assigned to treatment. These groups, referred to as *nonequivalent control groups* or *patched-up control groups*, help rule out specific rival hypotheses and decrease the plausibility of specific threats to internal validity.

6.8.1: Description and Rationale

One use of nonrandomly assigned subjects is to help rule out specific threats to validity, such as history, maturation, testing, and instrumentation. Such a group may be used when a no-treatment control group cannot be formed through random assignment. Although the purpose of this

group is exactly that of the randomly assigned no-treatment group mentioned earlier, there may be special interpretive problems that arise because of the way in which the group is formed. These groups are useful in helping to rule out threats to internal validity, but they may be weak for comparative purposes depending upon how they were formed.

Recall the quasi-experiment (The Pueblo Heart Study) designed to evaluate the impact of legislating in a city to have a smoke-free environment (Centers for Disease Control and Prevention, 2009). Pueblo, Colorado, had a smoke-free ordinance but evaluating its impact required groups to control for many potential threats to internal and construct validity (e.g., history, maturation; some other change that might have taken place such as rates of exercise, changes in health insurance). Two nearby cities did not have smoke-free ordinances and served as comparison cities. In relation to the present chapter, not quite equivalent control groups that could not be perfectly matched were used and in the case of this study made potential threats pretty implausible. Hospital rates for heart attacks changed from pre to post in Pueblo, but in the other two studies where the ordinance had not been implemented, the rates remained the same.

> In situations where random assignment is not possible (most situations in schools, cities, states, and federal), knowledge of methodology becomes especially important.

It is in such situations that one must begin with the concepts such as threats and what might interfere with drawing inferences. Control groups, conditions, or measures can be adopted to make threats just a little less plausible than the experimental manipulation one would like to evaluate.

6.8.2: Special Considerations

Nonequivalent control groups can vary widely and have to be evaluated on their individual merit. Their purpose is to reduce the plausibility that other influences (internal validity or construct validity) could explain the results. Because the group is not comprised randomly, the data may not be as persuasive as parallel data resulting from a randomly comprised control or comparison group. Yet, in any given case the absence of randomness may not be a fatal limitation. The question is whether some specific threat (e.g., selection *x* history or maturation) is as plausible as the interpretation the investigator wishes to place on the data. Although nonequivalent controls are less-than-perfect control groups, they can serve to tip the balance of plausibility among alternative interpretations of the data.

Another point to emphasize is about nonequivalent control groups. Individuals new to methodology or early in their careers may simply reject a study outright because randomness was not followed in devising a group. This view is arguable for the following reason. We do not worship practices (e.g., random assignment, between-group designs) in methodology; they are means to various ends, and one should invariably focus on the end.

> The "end" in this case is how plausible are threats to validity (all types) and whether anything in the study helps in making a potentially critical threat implausible. Random assignment can really help but is not perfect; nonequivalent control groups can help and are not perfect.

There is the added option in nonequivalent control groups of using matching techniques I have mentioned (e.g., propensity scores), and they too are not perfect. However, with matching a nonequivalent control group can actually be made increasingly equivalent to the other groups in the design. Such matching further reduces the plausibility that selection biases or special experiences or maturation of the nonequivalent or other groups are likely explanations of the findings.

> Nonequivalent control or nonrandomly assigned groups in clinical research usually address threats to internal validity. Groups might be added to provide useful information and to expand the conclusions that can be reached about the outcome and address construct validity too.

In treatment research, a valuable use of nonrandomly selected subjects is to compare the extent to which clients in the study are distinguished from their peers who have not been referred for treatment. By comparing individuals who have been identified as a treatment population with their peers who apparently are functioning with little or no problem, one can assess whether treatment has brought the clients within a "normal" range of behavior. The use of normative data to evaluate treatment is part of a larger area of evaluating the clinical importance of changes made in treatment.

6.9: Key Considerations in Group Selection

6.9 Identify some of the main deliberations while selecting a group

The previous discussion describes control and comparison groups that are likely to be of use in experimental research, in both true experiments and quasi-experiments in which the investigator is manipulating some experimental condition. There are no rules for deciding specifically what groups to include but there are guidelines that can help:

1. When developing a study, it is very helpful to ask oneself, "What do I need to control in this study?" This is not an open question without its own guidelines. Pull out the laminated wallet-sized copy of threats to validity (for the few readers who have not memorized them) and run down internal and construct validity threats in particular. Consider all of the threats, although many

may be able to be dismissed quickly. At the end of the study, one would usually want the threats to validity well controlled. So as you ask this question, you select a group and then re-ask, "Ok, if I included a control group like this (to be specified), would I cover (make less implausible) the scary threats to validity?" This first guide is simplified by noting, always be able to say precisely what your control group is designed to control for, i.e., do not merely say, "I used a control group."

2. As you design the study, ponder quite specifically what the results might look like. This can help decide what groups might be included before you actually begin the study. Initially, the "ideal" or expected results with respect to a particular hypothesis and prediction, if they can be specified, might be diagrammed; then more likely data patterns are considered.

   As variations of possible results are considered, the following question can be asked: "What other interpretations can account for this pattern of results?" The answer to that question is likely to lead to changes in the experimental groups or addition of control groups to narrow the alternative interpretations that can be provided.

   For example, you might include a special intervention group and a no-specific treatment control group. Sounds good, but ponder possible results. If both groups show improvement and equal improvement (no statistical difference), most of the threats to internal validity (e.g., history, maturation, testing, statistical regression) could explain this finding! Be careful. The last thing one wants to go to the trouble of conducting a study and only to have all of the threats to internal validity on the other side. More generally, pondering permutations of likely patterns or results and critical evaluation of rival interpretations of the findings are useful in generating additional comparison groups that are needed in a given study or bolstering the design by increasing the sample, to ensure a strong test of the major comparisons of interest.

3. Previous research also may dictate the essential control groups for a given investigation. For example, in the study of a particular treatment, it is not always necessary to use a no-treatment or wait-list control group. If there are consistent data that the absence of treatment has no effect, at least on the dependent measures of interest, these groups might be omitted. Of course, to justify exclusion of a no-treatment group, one would want convincing data about the likely changes over time without treatment. Relying on data from studies completed by other investigators at different research facilities might not provide an adequate basis to exclude a no-treatment group unless there is consensus that the problem is immutable without treatment. For example, "depressed clients" in one investigator's research may vary markedly from the "same" sample at another facility because of the different measures used in screening and the different locales. This is true even if depressed clients all meet criteria for major depression. They could still differ on the severity and duration of their depression and the presence of other disorders. On the other hand, within a research program, continued studies may reveal that no treatment or nonspecific-treatment groups lead to no change in the clients. In such a case, omitting these groups after several studies have been completed is somewhat more justifiable and also permits the investigator to move on to more sophisticated questions about the effects of treatment.

   As investigations build upon one another in a given areas of work, the research questions become increasingly refined, and there may be no need for some of the control groups used early in the research.

4. The selection of control and comparison groups may be dictated and also limited greatly by practical and ethical constraints. Practical issues such as procuring enough subjects with similar treatment problems, losing subjects assigned to control conditions for a protracted period, and related obstacles mentioned earlier may dictate the types of groups that can be used. Ethical constraints such as withholding treatment, delivering treatments that might not help or might even exacerbate the client's problem, deception about ineffective treatments, and similar issues also limit what can be done clinically. In the context of clinical samples, both practical and ethical issues may make it impossible to perform the comparisons that might be of greatest interest on theoretical grounds. There are other design options (single-case experimental designs) that are true experiments and that do not require control groups in the traditional way.

6.10: Evaluating Psychosocial Interventions

6.10 Assess how intervention research addresses the various research concerns

The use of various control and comparison groups isolated from an area of research is somewhat abstract. Also, the discussion does not convey the progression of research, which can be measured in the level of sophistication of the questions that are asked and the complexity of the conditions to which an experimental group is compared. Intervention research (e.g., psychosocial treatments, educational programs, medical procedures) nicely illustrates diverse control and comparison groups and the various research

Table 6.2: Intervention Evaluation Strategies to Develop and Identify Effective Interventions

Intervention Strategy	Question Asked	Basic Requirements
Intervention Package Strategy	Does the intervention lead to change (e.g., improvements after treatment)?	Intervention vs. no-intervention or waiting-list control group
Dismantling Intervention Strategy	What components are necessary, sufficient, or facilitative of change?	Two or more intervention groups. One receives the full intervention package; other groups receive that package minus one or more components.
Constructive Intervention Strategy	What components or other interventions can be added to enhance change?	Two or more intervention groups. One receives the full intervention package; other groups receive that package plus other components or the intervention.
Parametric Intervention Strategy	What changes can be made in the specific treatment and its delivery that will enhance change?	Two or more intervention groups that differ in one or more facets (e.g., duration, intensity).
Comparative Intervention Strategy	How effective is this intervention relative to other interventions for this clinical problem or intervention focus?	Two or more groups that receive different interventions.
Intervention Moderator Strategy	What patient, family, contextual, or other characteristics influence the direction or magnitude of change with this intervention?	One or more interventions but divided by levels of a predicted moderator (e.g., severity symptoms).
Intervention Mediator Strategy	What processes or constructs mediate the relation between the intervention and change?	One or more interventions with assessment on presumed processes that may be responsible, lead to, statistically account for change.

concerns the groups are designed to address. I will emphasize psychosocial interventions in the context of therapy to illustrate the control groups but also will draw from other areas to convey that the issues are not restricted to treatment of psychological problems.

> The goals of psychotherapy research are to identify effective treatments, to understand the underlying bases of therapeutic change, and to elaborate the client, therapist, and other factors on which treatment effects depend. The goals can be broken down into specific questions to build the knowledge base.

Major questions and the research strategies they reflect are noted in Table 6.2. A critical feature of the table is the control or comparison groups that are likely to be required to address the question of interest.

6.10.1: Intervention Package Strategy

The most basic question is to ask whether a particular treatment or treatment package is effective for a particular clinical problem. This question is asked by the treatment package strategy that evaluates the effects of a particular treatment as that treatment is ordinarily used.

> The notion of a "package" emphasizes that treatment may be multifaceted and includes many different components that could be delineated conceptually and operationally. The question addressed by this strategy is whether treatment produces therapeutic change.

To rule out threats to internal validity, a no-treatment or wait-list control condition is usually included in the design.

The package strategy with its no-treatment group remains common. As a brief example, a study was conducted to use music therapy for children (ages 10–12) with high aggressive behavior as reflected in scores on a parent-completed checklist (Choi, Lee, & Lim, 2008). Music therapy consisted of a therapist conducting two sessions a week for 15 weeks. The intervention included singing songs, making musical instruments, playing instruments such as the piano and hand bells, and more. Children assigned to the no-treatment control group received no intervention and were called regularly to be sure that they had not received some other form of treatment. The results indicated that those in the music therapy group were significantly more improved and different from nontreated children as reflected in reduced aggression and improved self-esteem. To the authors' credit, they noted that the results convey that the intervention led to change but one could not draw any conclusions about music therapy per se. The increased attention and contact of children in the program could readily account for the findings. Here is a good example where threats to internal validity were handled but we are left with a large construct validity problem. Was any music needed at all to achieve these effects?

Intervention package research can be a useful first step. If the package does not surpass a no-treatment group, one can go back to the drawing board (e.g., augment the package, change the intervention, change careers). If the package is different from no-treatment or a wait-list group, the next questions are about what of the treatment or why and how it worked.

Strictly speaking, evaluation of a treatment package only requires two groups, as in the example noted above. Random assignment of cases to groups and testing each group before and after treatment control the usual threats to internal validity. However, there has been considerable debate about the impact of nonspecific-treatment factors

and the effects they can exert on clinical dysfunction (e.g., Lambert & Ogles, 2013). Consequently, treatment package research is likely to include a group that serves as a nonspecific-treatment control condition or treatment as usual. Both of these latter groups of course require clients to come to the treatment and receive some active experience.

6.10.2: Dismantling Intervention Strategy

> The dismantling intervention strategy consists of analyzing the components of a given treatment package.

After a particular package has been shown to produce therapeutic change, research can begin to analyze the basis for change.

> To dismantle a treatment, individual components are eliminated or isolated from the treatment. Some clients may receive the entire treatment package, while other clients receive the package minus one or more components. Dismantling research can help identify the necessary and sufficient components of treatment.

An illustration of a dismantling focused on cognitive processing therapy for the treatment of posttraumatic stress disorder in female victims of interpersonal violence (past or present abuse or sexual assault) and who met criteria for posttraumatic stress disorder (Resnick et al., 2008). The therapy includes two main components:

- Cognitive processing
- Writing about the trauma (that is used as a basis for that processing)

In this study, the overall package with both of these components was provided to one group. The cognitive portion was provided to a second group (without the writing assignments); the writing portion was provided to a third group (without the cognitive processing). That is, the treatment was "dismantled" by seeing if the full package was needed. There were multiple measures (e.g., anxiety, depression, trauma) assessed before, during, and after the treatment. The major results: the treatment package, cognitive component only, and writing component only led to significant improvements and generally were no different in overall effectiveness. A study like this with similar results across the groups can raise many interesting substantive and conceptual questions (what are the mechanisms of action, do treatment components operate similarly, and do some individuals respond to one component better than others?). The results can also raise methodological questions (e.g., all three groups improving could be history, maturation, and repeated testing; also low power might preclude detecting what might be small differences). In any case, this is a good example of dismantling: compare a treatment package to other conditions in which components are separated. If there are three or more components, how and what to remove from the package and the number of comparison groups can become more complex.

6.10.3: Constructive Intervention Strategy

> The constructive intervention strategy refers to developing a treatment package by adding components to enhance outcome.

In this sense, the constructive treatment approach is the opposite of the dismantling strategy. A constructive treatment study begins with a treatment, which may consist of one or a few ingredients or a larger package. To that are added various ingredients to determine whether the effects can be enhanced. The strategy asks the question:

"What can be added to treatment to make it more effective?"

A special feature of this strategy is the combination of individual treatments. Thus, studies may combine conceptually quite different treatments, such as verbal psychotherapy and pharmacotherapy.

There is a keen interest in testing treatment combinations because the scope of impairment of many clinical problems (e.g., depression, antisocial personality) affects many different domains of functioning (e.g., symptoms, social and work relations). Also, many contextual influences on the individual (e.g., parents, spouses) may need to be integrated into treatment to help promote change or to reduce influences that may contribute to or sustain dysfunction in the client. For example, some families of patients diagnosed with schizophrenia are highly critical, hostile, and overinvolved, a set of characteristics that is referred to as *expressed emotion* and possibly mediated by heightened reactivity of patient in brain networks that process aversive social interactions (e.g., Rylands, McKie, Elliott, Deakin, & Tarrier, 2011). A single treatment (such as medication) that focuses on symptoms of schizophrenia (e.g., hallucinations, delusions) without attention to family interaction is limited. Several studies have shown that medication combined with a family-based component designed to address interpersonal communication significantly reduces relapse rates, compared to treatment without the family communication component (Pharoah, Mari, Rathbone, & Wong, 2010). The process through which these improvements may occur are not established, but one possibility is that family communication increases adherence to the medication and allows that intervention to work better.

Treatment combinations are often used in both clinical practice and research.

> The obvious view is that combining treatments may overcome the limits of any individual treatment and at the very worst would not hurt. There is an obvious way in

which combined treatments may be better. If the combined treatment addresses more facets of a problem (e.g., a broader range of symptoms or range of factors that maintain the problem), or if the treatments act in concert (e.g., produce some interactive effect), then they are likely to be more effective than the individual components.

Combinations of some medications (e.g., HIV) operate in this way and can be shown to be more effective than the constituent medications given by themselves. It is not always better or even likely to be better to provide combined treatments, counter to common assumptions.

Combined treatments come at a price. If treatment is medication, then the number of side effects or problems of adherence (take two or more medications) raise obstacles.

If treatment is psychotherapy and the duration of treatment is fixed (e.g., only a certain number of therapy sessions), then squeezing in two (or more) treatments into this period may dilute the individual components and their effects. Clients will not remain in therapy forever, of course, while the therapist provides her or his special brand of combined treatments. In studies that combine treatments, it is obviously important to include as a comparison condition a group in which the most powerful constituent treatment or each of the treatments that comprise the package is evaluated alone.

6.10.4: Parametric Intervention Strategy

The parametric intervention strategy refers to altering specific aspects of treatment to determine how to maximize therapeutic change.

Dimensions or parameters are altered to find the optimal manner of administering the treatment. These dimensions are not new ingredients added to the treatment (e.g., as in the constructive strategy) but variations within the technique to maximize change.

Increases in duration of treatment or variations in how material is presented are samples of the parametric strategy.

A basic parameter of treatment is duration (number of sessions or amount of time for a given session) or a range of tasks, stimulus materials, opportunities for practice, feedback, and reinforcement, depending on variables within a given intervention that might be manipulated. As an illustration, a recent study compared two groups that allowed variation of a parameter, in this case dose, of treatment (Molenaar et al., 2011). Adults with depression were randomly assigned to one of two groups: 16 sessions of psychotherapy combined with pharmacotherapy or 8 sessions of the same treatment. The outcome was alleviation of depression and improving social functioning. The findings—both groups were equally effective and hence in this study the amount of therapy made no difference.

Parametric studies can study all sorts of other variables than just dose. Varying combinations of components of treatment and the ordering of different components are two examples. The key that defines parametric is variation of some facet of a given treatment that is not dismantling components of the intervention.

6.11: Evaluating Additional Psychosocial Interventions

6.11 Evaluate three additional psychosocial strategies that can be used to develop effective interventions

The additional psychosocial interventions that we will evaluate in the context of therapy to illustrate the control groups are comparative intervention strategy, intervention moderator strategy, and intervention mediator strategy.

6.11.1: Comparative Intervention Strategy

The comparative intervention strategy contrasts two or more treatments and addresses the question of which treatment is better (best) for a particular clinical problem.

Comparative studies attract wide attention not only because they address an important clinical issue but also because they often contrast conceptually competing interventions. Historically, the comparative studies have been more than scientific scrutiny of interventions; they have been battlegrounds for treatments that were quite different in their foci and conceptual views.

In keeping with this battleground view, comparative outcome studies included psychoanalysis versus behavior therapy, cognitive therapy versus medication, family therapy versus individual therapy, and many others.

Multiple factors have kept comparative studies as a central focus but the tenor has changed from pitting one treatment against another to make a broad conceptual point. Among the reasons, more efforts have been made to integrate diverse conceptual views. Also, treatments that were once argued as "pure versions of something" often have components (e.g., therapeutic alliance, homework assignments) of other treatments to which they might be contrasted. Even some conceptual views once considered diametrically opposed include elements (e.g., acceptance, mindfulness) that are common to many diverse techniques. For example, some behavioral techniques (e.g., dialectical behavior therapy, parent–child interaction therapy) give considerable attention to relationship issues and that attention overlaps greatly with traditional talk therapies. In years

long gone, this overlap would have been nonexistent, minimized, or denied. The net has made sharp contrasts of very different treatments less salient. There remains interest in discovering what treatments work the best but relatively few direct head-to-head comparisons are made.

The resurgence of interest in comparative research comes from the U.S. Government and the call for comparative effectiveness research (CER) in health care more generally (United States Department of Health and Human Services, 2009b; National Institutes of Health, 2010, 2013b). Impetus derives in part from the fact that patient care includes many health-related areas. Often the most effective or evidence-based interventions are not used and when they are used, we may not know which among alternatives are most effective. The U.S. government has set as a national priority evaluation of the relative effective of available procedures, and this includes biological and psychological interventions. Also part of this is a huge area referred to as complementary and alternative interventions, and these include interventions outside of conventional care (e.g., herbs, spas, acupuncture, meditation, spiritual practices). Complementary and alternative interventions are widely used through the United States and the world (e.g., Frass et al., 2012; Su & Li, 2011). In the United States, for example, approximately 40% of adults use some complementary and alternative intervention for their personal care (Kaptchuk & Phillips, 2011). One can see more broadly from a social and policy perspective, it would be important to identify what the treatments are for various physical and psychological sources of impairment and which among these are the most effective. This latter focus is of course the comparative intervention strategy.

A recent comparative study is of interest because it raises old and new battleground issues but also is quite contemporary. This study compared cognitive therapy with psychodynamic supportive psychotherapy to treatment major depression in adults (Driessen et al., 2013). The study was a randomized controlled trial (RCT) conducted with a large sample (N = 341) and carried out at three different sites. Adults who met psychiatric diagnostic criteria for major depression were assigned randomly to receive cognitive therapy, arguably the standard and most well-studied psychosocial intervention for depression, and psychodynamic supportive therapy. Individuals with especially elevated scores on one of the measures received medication as well as the treatment to which they were assigned. Treatments were administered for 22 weeks (16 sessions), and there was a 1-year follow-up.

The results indicated that the treatments were equally effective with 24% and 21% of the patients showing recovery (defined by stringently low scores on one of the measures of depression) for cognitive therapy and psychodynamic therapy, respectively. The study was well done in many ways (final sample > 230 subjects with data for the primary data analysis, a much larger sample that most therapy studies, therapist training and supervision). Invariably there are some methodological points with impact difficult to evaluate (no control group precludes evaluation of changes due to the usual threats to internal validity and the common factors threats to construct validity, also assessors were not completely blinded in relation to the outcome assessment). Support for the null hypothesis (no difference) in the model of quantitative research invariably raises questions (because no difference can be due to so many things). All that said, the fact is the study showed both groups improved and did so equally. In the process, this nicely illustrates a comparative outcome study.

The study is important for other reasons; psychodynamic therapy has not enjoyed the same degree of attention in rigorous studies, and so inclusion in a major multisite study and showing effects equal to cognitive therapy are notable. Also the results convey another critical point. Although the treatments were equally effective, arguably they were not very effective. Only 21–24% of the patients met criteria for recovery indicating that more, different, or better treatments for depression are still needed.

6.11.2: Intervention Moderator Strategy

The previous strategies emphasize the technique as a major source of influence in treatment outcome and search for main effects of treatment, i.e., that treatment is better or worse for all of the individuals as a group. Yet it is much more likely that the effectiveness of treatments varies as a function of multiple other variables related to individuals, contexts in which they live, and so many other factors. Those other variables or factors are called moderators. We have discussed and illustrated moderators previously. As noted then, moderators are variables that influence the magnitude of effect or the direction of effects of some other condition or variable (e.g., in this case treatment).

In the usual conceptualization of this strategy in relation to treatment, characteristics of the clients or therapists or the treatment process (therapeutic alliance) are the usual focus. The strategy would be implemented by selecting clients and/or therapists on the basis of specific characteristics. When clients or therapists are classified according to a particular selection variable, the question is whether treatment is more or less effective with certain kinds of participants. For example, questions of this strategy might ask if treatment is more effective with younger versus older clients, or with certain subtypes of problems (e.g., of depression) rather than with other subtypes.

As discussed previously, one is guided by theory or informed hypotheses to select moderators for investigation. Clinical common sense might be a guide as well. For we know that children, adolescents, and adults who meet

criteria for one psychiatric disorder are likely to meet criteria for one or more other disorders, a phenomenon referred to as *comorbidity* (e.g., Byers, Yaffe, Covinsky, Friedman, & Bruce 2010; Wichstrøm et al., 2012). When one is evaluating treatment, perhaps the effectiveness will depend on (be moderated by) whether participants meet criteria for another disorder, what those other disorders are, and their severity. That is, comorbidity (meeting diagnostic for more than one disorder) may moderate treatment outcome. Comorbidity is one client characteristic that would be a reasonable focus for the client and therapist variation treatment evaluation strategy. This of course is one possible moderator.

> The overall goal of this evaluation strategy is to examine factors that may moderate treatment effects, i.e., whether attributes of the client, therapist, or context contribute to outcome. One goal of studying moderators is to do better triage, i.e., directing people to treatments from which they are likely to profit and away from treatments that are likely to fail.

This is part of the rationale for "personalized medicine," namely, identifying moderators that direct what treatments are provided.[1] Moderator research in clinical psychology has not helped at this point in directing patients to treatments by showing because the work rarely shows that a particular treatment will not be very effective with one type of problem or client but another treatment will be. Also, moderators may affect the magnitude of relations rather than stark conclusions about a particular treatment working or not working with a client group.

For example, in my own work we have found barriers to treatment participation as a moderator of treatment outcome among families with children referred for aggressive and antisocial behavior. Barriers refer to parental perceptions of stressors related to participating in treatment (e.g., seeing treatment as demanding and not well suited to their child). Families who perceive greater barriers to treatment show less therapeutic change than families who perceive fewer variables (Kazdin & Wassell, 2000; Kazdin & Whitley, 2006). These effects are evident while controlling for other potential confounding variables (e.g., family stress outside of the context of treatment, parent psychopathology, severity of child dysfunction). Interestingly, parents with high barriers still show improvements, so treatment does not fail with them but clearly the magnitude of change is moderated by barriers to participation.

6.11.3: More Information on Intervention Moderator Strategy

It is helpful to know when moderators do not seem to make a difference. For example, a review of multiple studies for cognitive therapy for the treatment of obsessive compulsive disorder in children and adults indicated that several likely moderators (e.g., duration of symptoms, age of onset, comorbidity) did not make much difference in treatment outcome (Olatunji, Davis, Powers, & Smits, 2013). This information is useful to know by conveying that an effective treatment may not be equally applicable across clinical samples and age groups.

Research on moderators can be enlightening. Rather than main effects of treatment (for all individuals in the experimental or treatment group), the question focuses on interactions (whether some individuals respond better to treatment than others or whether some individuals respond to one form of treatment whereas other individuals respond better to another form of treatment). Different types of variables (treatment, subject, contextual) are combined. Although the usual control conditions might be used, the question usually focuses on comparison groups that are composed of combinations of treatment and subject characteristics.

There are a couple of important limitations in treatment research on moderators:

1. It is likely that multiple moderators are involved. Most research on treatment moderators plods along with one moderator at a time.
2. Once a moderator is studied, why and how it works is rarely pursued. Thus, we have a description of isolated moderators without little idea of how to operate. This limits our ability to use the information to make treatments more effective. That is, if we know what was going on and the processes through which a moderator achieved its effect, we might be able to make changes in the moderator or accommodations in treatment to improve outcomes.
3. Studying one moderator at a time is an enormous limitation in understanding how and for whom treatment works. It is likely that multiple moderators are involved in contributing to change. Recently, methods for integrating and combining multiple moderators have been elaborated (Kraemer, 2013). Individual moderators tend to be weak in how they predict outcome (e.g., effect size) and may not even emerge as statistically significant. Yet, multiple moderators can be combined and with that combination meaningful effects of moderator x treatment interactions emerge that otherwise would not be evident (e.g., Frank et al., 2011; Wallace, Frank, & Kraemer, 2013).

All these points notwithstanding, the search for moderators represents a more sophisticated approach to treatment evaluation than the search from main effects alone.

6.11.4: Intervention Mediator/ Mechanism Strategy

The previously noted strategies emphasize outcome questions or the impact of variations of the intervention on clients at the end of or subsequent to treatment. The treatment

mediator strategy addresses questions pertaining to how change comes about. What processes unfold that are responsible for improvement? As we have discussed and illustrated mediators (and mechanisms) previously, we can be brief here.

Much of the research using the treatment mediator strategy has looked at a particular construct (e.g., changes in specific cognitions) and how it relates to treatment outcome (Kazdin, 2007). The view is that the treatment technique achieves its effects (therapeutic change) through altering specific cognitions (mediator). When such findings are established, this does not mean change in cognitions caused the change but rather there is a special statistical association. That association usually means that some intervention (treatment) led to change (outcome) and that the change in the outcome depended on (was associated statistically with) some intervening process variable (changes in cognitions). Furthermore, if these cognitions did not change, the outcome was not likely to occur. We cannot say that cognitions caused the change. It could be that cognitions are correlated with some other influence. Even so, research on mediators can move our knowledge forward by ruling out both influences that not likely to be involved and influences that are. This dual effect of ruling in and ruling out likely mediators is nicely illustrated in a study that looked at several mediators.

This study was an RCT of treatment of college student drinkers (whose treatment was mandated) and who received motivational enhancement therapy (LaChance, Feldstein Ewing, Bryan, & Hutchison, 2009). Motivational enhancement therapy is an intervention often used with addictions. Over usually a brief number of sessions, clients are provided feedback for their behavior and encouraged to better understand their motivations and improve self-control. In this study, five mediators were examined to explain the basis for therapeutic change:

- Readiness to change
- Self-efficacy
- Perceived risk
- Norm estimates (what others are doing)
- Positive drinking expectations

Only self-efficacy served as a mediator. The extent to which individuals gained a sense of agency or control was associated with improved outcome. This is potentially quite informative. Among the next steps, for example, might be to manipulate self-agency directly to see if it is causally involved in the change process and if treatment outcome effects could be enhanced. As for a study designed to evaluate mediators, this is exemplary because of the investigation of several in the same project. Assessing multiple mediators is not only efficient in evaluating mediators, but also raises the possibility of identifying subgroups and evaluating whether different mediators are involved (moderated mediation).

> Many therapy studies focusing on mediation have turned to neuroimaging. The purpose is to look at therapeutic change and how those changes are related to changes in brain activity, often in areas of the brain already implicated in the disorder based on prior research.

Among the goals is to identify whether treatment leads to changes in brain activity and brings that activity closer to normative levels as defined by individuals without the clinical dysfunction in the treated sample (e.g., Frewen, Dozois, & Lanius, 2008; Quidé, Witteveen, El-Hage, Veltman, & Olff, 2012).

As with other mediators, changes in brain structure, function, and activity do not establish causal links, but they home in on locales where one could search for precisely what and how changes come about (see Kazdin, 2014). Also, brain activity can lead to finer-grained hypotheses (e.g., hormonal, neurotransmitter, synapse) that can encompass and draw on data that focus on the development and progression of disorders outside of the context of treatment research.

As a general strategy, we want to know why a particular intervention or experimental manipulation works so that treatment mediation strategy is an effort to move further toward that.

> Mediation can move us closer to understanding specific process that might be involved.

Further research can follow up on mediation studies in an effort to identify if specific processes if altered (enhanced or blocked) can influence the outcome. This is an excellent instance in which human and nonhuman animal studies often are involved in moving back and forth from laboratory studies in critical processes to the clinic with strategies to improve patient care (Kazdin, 2014).

6.11.5: General Comments

The strategies noted previously reflect questions frequently addressed in current intervention research (treatment, prevention, education, rehabilitation). The questions posed by the strategies reflect a range of issues required to understand fully how an intervention operates and can be applied to achieve optimal effects. The treatment package strategy is an initial approach followed by the various analytic strategies based on:

- Dismantling Research
- Constructive Research
- Parametric Research

The comparative strategy probably warrants attention after prior work has been conducted that not only

indicates the efficacy of individual techniques but also shows how the techniques can be administered to increase their efficacy. Frequently, comparative studies are conducted early in the development of a treatment and possibly before the individual techniques have been well developed to warrant such a test. A high degree of operationalization is needed to investigate dismantling, constructive, and parametric questions. In each case, specific components or ingredients of therapy have to be sufficiently well specified to be withdrawn, added, or varied in an overall treatment package.

The progression requires a broad range of control and comparison groups that vary critical facets of treatment. The usual control conditions (no-treatment, nonspecific-treatment control) may continue to play a role. However, the interest in evaluating change over time without treatment or factors common to treatment gives way to more pointed questions about specific facets of treatment that account for or contribute to change. Comparison groups are aimed to allow increasingly specific statements related to construct validity, i.e., what aspects of the intervention account for the findings?

Not all of the questions one might ask of a given intervention are addressed by the strategies we have discussed. For example, once an intervention is effective with a specific disorder or domain of functioning, it is natural to extend the intervention to see if related domains are also altered. This might be considered beginning the treatment package strategy anew but just applying this to another problem.

For example, medication for the treatment of cigarette smoking recently was also shown to be effective for alcohol dependence (Litten et al., 2013). Cigarette smoking and alcohol dependence often go together and share biological underpinnings (e.g., receptors in the brain) and the medication that was studied (Verenicline [marketed under the name Chantix]) works on those receptors. In any case, the intervention strategy is to see if a treatment effective for one problem can be effective for another. Test of generality of the impact of a treatment might focus on different types of disorders, clients, and settings. One might consider these variations of the treatment moderator strategy, namely, does the effectiveness of an intervention vary as a function of other ways (to whom, how) in which it is applied?

Summary and Conclusions: Control and Comparison Groups

Control groups rule out or weaken rival hypotheses or alternative explanations of the results. The control group appropriate for an experiment depends upon precisely what the investigator is interested in concluding at the end of the investigation. Hence all, or even most, of the available control groups cannot be specified in an abstract discussion of methodology. Nevertheless, treatment research often includes several specific control procedures that address questions of widespread interest.

The no-treatment control group includes subjects who do not receive treatment. This group controls for such effects as history, maturation, testing, regression, and similar threats, at least if the group is formed through random assignment. The wait-list control group is a variation of the no-treatment group. While the experimental subjects receive treatment, wait-list control subjects do not. After treatment of the experimental subjects is complete, wait-list control subjects are reassessed and then receive treatment. A no-contact control group may be included in the design to evaluate the effects of participating in or having "contact" with a treatment program. Individuals selected for this group usually do not know that they are participating in a treatment investigation. Hence their functioning must be assessed under the guise of some other purpose than a treatment investigation. More commonly now in light of self-help treatments, no-contact is less of a control group than a way of administering treatment with little or no contact with a therapist.

A nonspecific-treatment control group consists of a group that engages in all of the accouterments of treatments such as receiving a rationale about their problem, meeting with a therapist, attending treatment sessions, and engaging in procedures alleged to be therapeutic. Actually, the purpose is to provide those ingredients that could lead to change but are not central to the intervention that is being evaluated. This control condition allows one to address of whether the effects of veridical treatment are merely due to its nonspecific-treatment components. This is a critical construct validity issue.

Treatment as usual consists of the usual, routine, and standard care treatment that is provided for a problem at a particular clinic or other setting. Clients assigned to this treatment receive a veridical intervention (unlike a nonspecific-treatment control condition), and many of the factors common to most treatments are controlled. Few objections arise from therapists and clients regarding the use of routine care as a comparison condition. From a methodological standpoint, a difficulty with treatment as usual is

that it is usually unstructured and unspecified, varies from clinic to clinic and therapist to therapist, and therefore is not replicable. This might be remedied by specifying what was done and trying to achieve consistency among therapists, but those efforts would make this treatment as no-so-usual.

A yoked group controls for variations across groups that may arise over the course of the experiment. Implementing treatment procedures may involve factors inherent in but not relevant to the independent variables of interest to the investigator. Yoking refers a procedure that equalizes the extraneous variables across groups by matching or pairing subjects in the control groups (or one of the control groups) with subjects in an experimental group and using information obtained from the experimental subject to decide the conditions to which the control subject will be exposed.

Nonequivalent control groups refer to a category of groups that is characterized by selection of subjects who are not part of random assignment. These groups are added to the design to address specific threats to validity (usually internal validity such as history or maturation) that are not handled in the usual way (e.g., random assignment to experimental and no-treatment control groups). A nonequivalent control group, by virtue of its selection, imperfectly controls these threats but still strengthens the plausibility of the conclusions that can be drawn.

The addition of control and comparison groups to experimental designs usually addresses threats to internal and construct validity and hence adds precision to the conclusions that can be reached. The progression of research and the different control and comparisons groups that are used were illustrated in the context of psychotherapy research. Many different treatment evaluation strategies were discussed to convey various control and comparison groups and questions that do not require control conditions in the usual sense.

Many designs that are used in psychology do not involve experiments where there is a manipulation or assignment to conditions. In these designs, various samples (e.g., individuals exposed to domestic violence vs. not; individuals with a particular disorder vs. another) are evaluated and compared. The goals of such research include developing and understanding of various conditions and their impact. These goals too require control and comparison conditions. Observational designs are the focus of the next chapter along with the conditions required to address threats to validity.

Critical Thinking Questions

1. An experiment may show that a treatment (cognitive behavior therapy) is better than no-treatment and controls for all of the threats to internal validity, but is likely to have a construct validity problem. What is that problem?
2. Developing an attention-placebo control group has special challenges. What are they?
3. What are some of the strengths and weaknesses of using treatment as usual as a control group?

Chapter 6 Quiz: Control and Comparison Groups

Chapter 7

Case-Control and Cohort Designs

Learning Objectives

- **7.1** Explain how observational research plays an important role in certain fields like psychology
- **7.2** Define case-control designs
- **7.3** Compare case-control designs with cohort designs
- **7.4** Analyze how prediction, classification, and selection are ways of referring to some outcome
- **7.5** Identify the specific issues that the researcher needs to be aware of at the research design stage
- **7.6** Express the importance of proper specification of the construct due to its impact on the findings
- **7.7** Recognize the importance of selecting the right group in research
- **7.8** Determine how incorrect reporting of the predictor and the outcome leads to incorrect findings
- **7.9** Report the utilities of case-controlled designs over experimentally studied ones

Up to this point, we have focused primarily on true-experimental designs in which subjects are randomly assigned conditions and the variables of interest are manipulated experimentally by the investigator. We also covered many of the control and comparison groups that these designs often include. In much of clinical research, subject characteristics and other variables are not manipulated directly by the investigator. Rather, the variables are "manipulated by nature" and the investigator evaluates the impact of these variables through selecting persons for study who have the characteristic of interest. Such studies are sometimes referred to as *observational research to convey that the role of the investigator is to observe (assess) different characteristics and their associations, rather than to intervene experimentally*. Although observational research can identify many patterns of association (correlates) and can describe the nature of various characteristics (e.g., disorders), the goals are to develop and test theories and to understand causal relations in much the same way as experimental research.

There are many options for observational research. This chapter considers major design strategies, with an emphasis on those that are more commonly used in psychological research. In each design strategy, the central characteristics are the study of intact groups (e.g., no random assignment) and examination of variables and influences that the investigator usually cannot manipulate directly. The designs have the same goal as experimental designs, namely, to make implausible various threats to validity. Innovative methodological thinking and practices often are called on because the investigator does not have the luxury of an experiment where manipulation of conditions and assignment of subjects are controlled.

7.1: Critical Role of Observational Research: Overview

7.1 Explain how observational research plays an important role in certain fields like psychology

Designs in which intact groups are studied concurrently or over time are not presented very often in teaching research design in psychology. For one reason, there is a

strong experimental tradition (i.e., true experiments) within psychology in which direct manipulation is accorded somewhat higher status than so called "correlational" research. Well recognized is that one of the best ways to demonstrate a causal relation is to manipulate something directly and see if the predicted outcome changes. In clinical research, that is why the randomized controlled trial has special status in demonstrating the effectiveness of an intervention (e.g., psychological treatment, surgery, medication). Yet, as we discuss in this chapter, observational research has special status too in identifying relations and findings that could never be evaluated experimentally.

Another reason the designs are not emphasized in psychology may stem from their primary association with other disciplines. For example, these designs rule in epidemiology and public health where intact groups (populations at risk, with disease) are routinely studied. The designs and methods of data analyses are specialties of their own. There is barely enough time to teach some experimental designs and some statistics in psychology, yet draw on the methodological advances of other disciplines areas. Yet the key components of methodology (research design, data evaluation) span many areas of science, and the increase interdisciplinary collaborative nature of research has help diffuse methodologies across boundaries.

It is important to dispel quickly a traditional psychology view, perhaps not as readily voiced today, that observational research has secondary status and takes a back seat to true experiments. Sciences in general are more in the public view, and we see regularly that many if not most scientific fields (e.g., astronomy, archeology, meteorology, volcanology [volcanos], seismology [earthquakes], and of course my favorite, pomology [study of fruits]) rely heavily on observations of different conditions rather than experimental manipulation of the subject matter. Now we can predict catastrophic weather conditions fairly well and also identify planets from other galaxies (called exoplanets) that might be habitable and provide new places to start up fast-food franchises. Few scientists or citizens complain that both the weather predictions and exoplanets emerge from observational data alone.

In psychology, and perhaps especially clinical, counseling, school, and educational psychology, observational research plays a special role for several reasons. They are:

1. Core questions of interest do not permit experimental manipulation. For example, even as debate continues about how to diagnose mental disorders, we study them. Indeed, our studies will shed light on domains that will eventually improve diagnosis. We study these disorders all of the time to shed light on the risk, onset, etiologies, and course. These are primarily observational studies of different patient groups (e.g., individuals with and without depression).
2. We are very interested in individuals with special experiences due to exposure (e.g., to trauma, war, domestic violence, prenatal cigarette smoking) or to deprivation (early loss of parents, malnutrition).

   There is a long tradition in psychological research in studying special groups (e.g., individuals who are first born among siblings, criminals, octogenarians, Nobel laureates, methodologists). These foci require studying or observing intact groups. In each of these areas, research is designed to address a host of questions, such as what are the past, present, and future characteristics of such individuals? What factors predict who will show the outcome of interest? What are the causes of the outcome? And even, what may be done to prevent the outcome?

   Obviously, one cannot assign individuals to experience one condition versus another (e.g., receiving harsh vs. mellower child rearing; receiving vs. not receiving a Nobel Prize, being or not being exposed to animated methodology videos prenatally). However, individuals with these varying characteristics can be identified and studied.
3. The influence of other disciplines on clinical research has expanded the design strategies that are used within psychology. Epidemiology and public health have impact on clinical psychology, psychiatry, and related disciplines (e.g., health psychology, psychiatric epidemiology). For example, the vast majority of public health studies on the factors leading to diseases (e.g., AIDS, heart disease, and various forms of cancer) have come from observational, rather than experimental studies. Psychology studies these disorders too and uses the same observational research methods. From observational research, we have learned about multiple influences on diseases (morbidity) and death (mortality), the relative weight of various influences, and whether some influences are likely to play a causal role or are piggybacking on some other variable. The designs can be very powerful indeed. Often we can take the findings from observational research and move them back to the laboratory with nonhuman animal models to see if there are causal relations.
4. Models in science have evolved in ways that also accord greater value to observational designs. Experimental research, as powerful as it is, is often restricted to the manipulation of one or two variables at a time. Isolation of variables is a key advantage of experimentation to understand how variables operate. However, in many areas of science (e.g., physiology, meteorology, economics), including psychology, we know that there are multiple variables that may influence a phenomenon of interest and that these variables may be related in dynamic

(constantly changing), interactive, and reciprocal ways. Often we want to study systems, large units with interacting processes. Observational studies can take into account (e.g., with statistical and math models) multiple variables, study them over time, and examine the influences of variables on each other.

5. Data-analytic techniques have advanced over the past decades that can strengthen the inferences drawn from observational research (e.g., Little, 2013). Diverse methods of analysis (e.g., path analysis, structural equation modeling, and hierarchical linear regression) have evolved and are increasingly familiar; other methods widely used in other disciplines (e.g., logistic analysis, survival analysis, time-series analysis) are used increasingly in clinical, counseling, and educational psychology. The net effect is to provide better means of drawing inferences from longitudinal data and the direction and type of influence that one variable exerts on another. The findings have provided information that could not be obtained from experimental research. This means one can tease out, separate, and evaluate the influence of factors that may be confounded with the group status and progress to increasingly nuanced questions.

7.1.1: More Information on the Critical Role of Observational Research

Recall simple relations from observational studies with group comparisons. For example, cigarette smokers (one group) have higher rates of heart disease, lung cancer, and early death compared with nonsmokers. In that observational study, we want to control those variables that might be confounded (e.g., cigarette smokers drink more alcohol and exercise less) and introduce a construct validity problem (e.g., is it smoking or some other construct)? And now observational research moves to nuanced questions—what are characteristics of cigarette smokers who live long healthy lives or those individuals who never smoke and die of lung cancer? These are enormously important questions that can be elaborated by observational studies and data-analytic techniques to reveal otherwise obscure relations.

> A key theme of this text is the need and role for multiple methodologies. To that end, there is no need to pit true-experiments and observational research against each other. In science, diverse methods are needed and they are complementary.

For example, we learned decades ago that exposure to low levels of lead (in water, air, and diet) among children is associated with hyperactivity, deficits in neuropsychological functioning (e.g., verbal, spatial ability), distractibility, lower IQ, and overall reduced school functioning, and these effects continue to be evident several years later (Needleman, Schell, Bellinger, Leviton, & Alldred, 1990). This and many other studies were observational in nature and compared children exposed or not exposed to lead and those exposed in varying degrees. Observational studies of humans were followed by true-experiments with nonhuman animals (in rats, monkeys) that showed exposure to lead influenced brain activity and structure (e.g., neurotransmitter activity, complexity in dendrite formation, inhibition of the formation of synapses) and hence elaborated how learning and performance are inhibited. The point in mentioning a slice of this large area of research is to illustrate a back and forth of observational research and experimental studies as well as human and nonhuman animal studies. All are needed to obtain the understanding we wish and in the case of this example (with low levels of lead) they have had enormous implications for prevention of physical and mental health problems in children (see Centers for Disease Control and Prevention, 2012a).

7.2: Case-Control Designs

7.2 Define case-control designs

There are many options for observational research, and we begin with the most familiar and basic. (Table 7.1 includes the specific designs we will cover and provides a useful summary guide.)

Table 7.1: Selected Observational Designs: Summary of Key Characteristics

Design	Summary Characteristics
1. Case-Control Designs	Investigation of a characteristic of interest by forming groups who vary on that characteristic and studying other current or past features of those groups
a. Cross-sectional case-control design	Identify cases (individuals with the characteristic of interest) and controls (without the characteristic) and evaluate other characteristics currently, i.e., evident at this point in time
b. Retrospective case-control design	Identify cases (individuals with characteristic of interest) and controls (without the characteristic) and evaluate other characteristics in their past in an effort to identify antecedents of the current outcome
2. Cohort Designs	Investigation of intact group(s) over time but prospectively (longitudinally)
a. Single-group cohort design	Identify subjects who meet a particular criterion (e.g., exposed to an event such as a national disaster, or born in a given year, or with some specific characteristic) and follow that group prospectively to assess an outcome of interest (e.g., onset of a disorder). Birth-cohort design is a special case of this design.
b. Multigroup cohort design	Two or more groups are identified who meet a particular criterion and are followed prospectively to assess an outcome of interest (e.g., onset of a disorder)
c. Accelerated, multi-cohort longitudinal design	Two or more groups are selected that vary in age (different cohorts) and who are followed prospectively. The design is "accelerated" because a longer period of development is covered by selecting cohorts at different periods and following them.

Case-control designs refer to strategies in which the investigator studies the characteristic of interest by forming groups of individuals who vary on that characteristic and studying current or past features of the groups.

The key characteristic is in identifying groups who vary in the outcome (criterion) of interest, i.e., have the "problem" or characteristic that the investigator wishes to elaborate. Case-control design is the term used extensively in epidemiology and public health where "case" typically means someone who has the disease or condition (e.g., heart disease, high blood pressure) that is to be studied. For psychology, "case" merely refers to individuals with the characteristic of interest.

In the most basic, two-group version, the investigator compares subjects who show the characteristic (cases) with individuals who do not (controls). The independent variable is the characteristic or criterion that served as the basis for selection and may reflect a particular experience (e.g., being victimized, exposure to a particular parenting style) or status (e.g., being first born, widowed, divorced). The investigator compares the two groups on the measures of interest and then interprets the differences to reflect a critical facet of the problem. Two major variations of the designs are worth distinguishing, based on the time perspective in which the groups are studied.

7.2.1: Cross-Sectional Design

In a cross-sectional, case-control design, the most commonly used version in psychology, subjects (cases and controls) are selected and assessed in relation to current characteristics. This is distinguished from studies that are designed to evaluate events or experiences that happened in the past (retrospective studies) or that will happen in the future (prospective studies).

The goal of a cross-sectional, case-control study is to examine factors that are associated with a particular characteristic of interest at a current point in time.

The study can describe and explore characteristics of interest (e.g., what are peer and family relations like of young women who have poor vs. good body image) or test theoretical propositions or conceptual models (e.g., first- and second-born children might be compared to test a hypothesis about different patterns of attachment in their current adult relations; children who grew up in bilingual homes vs. those who did not on subsequent implicit attitudes toward other ethnic groups; ethnic differences in perceived barriers to obtaining mental health services). Obviously, the questions are endless.

In the usual case, the investigator begins with hypotheses about how various groups will differ, perhaps based on a theoretical perspective. The subjects are identified and assessed on multiple characteristics beyond those used to delineate their status as cases or controls. Because all of the measures are obtained at the same point in time, the results are correlational, i.e., one cannot know from the study whether the outcome preceded or was caused by a particular characteristic. (There are some exceptions where a characteristic such as sex or ethnic identity may be assumed to antedate an outcome of interest such as onset of a disorder.)

Cross-sectional designs are useful for identifying correlates and associated features, and these findings may be quite informative and significant.

For example, the investigator may wish to test whether depressed mothers interact differently with their children (e.g., infants, toddlers) when compared with nondepressed mothers. Mothers are identified and assessed on a measure (or two) of depression and classified as depressed (cases) or not (controls); they are then brought into the laboratory or observed at home to assess how they interact with their children. Several studies with this focus have shown that depressed mothers, compared with nondepressed controls, display decreased attention, affection, and vocal behavior, are less expressive (flatter affect) and show more anger, negativism, and hostility in their interactions (e.g., Conroy, Marks, Schacht, Davies, & Moran, 2010; Field, 2010). This work has very important implications regarding early child development, patterns of emotional attachment of parents to children, and the likely risk that children may have for later dysfunction.

Cross-sectional designs are commonly used and have generated provocative findings theories, and further research. For example, from such studies we have learned that:

- Individuals who are depressed are likely to show a set of negative cognitions (e.g., helplessness, hopelessness) compared with nondepressed controls.
- Children whose parents survived the holocaust experience significantly greater psychological dysfunction than matched controls whose parents have no such experience.
- Children who are depressed compared with those who are not have significant impairment in peer relations and school functioning.
- Girls who mature early (in relation to their peers) are more likely to have low self-esteem than those who mature later, to mention a random (well not entirely random) list of fascinating findings.

Many examples mentioned previously related to the health benefits of drinking wine, participating in religion, and not being depressed after a heart attack were based on case-control studies. Findings that compare intact groups are very useful in generating theory and concrete hypotheses to analyze further the reasons for these relations and

the conditions under which they do and do not operate. Indeed, many case-control studies lead to experimental research (e.g., studies with nonhuman animals) to test directly some characteristic expected to play a causal role that may be suggested by a case-control study.

7.2.2: Retrospective Design

In a retrospective, case-control design, the goal is to draw inferences about some antecedent condition that has resulted in or is associated with the outcome. This design represents an explicit effort to identify the time line between possible causes or antecedents (risk factors) and a subsequent outcome of interest. Subjects are identified who already show the outcome of interest (cases) and compared with those who do not show the outcome (controls). So far this is just like the cross-sectional case-control design.

> The retrospective design includes measures that are designed to elaborate the past of the individuals in each of the groups. Essentially, the design says you are this way now (case or now); now tell us about some things in your past (antecedents).

As an example, a retrospective case-control design was used to evaluate the relationship of breastfeeding and attention deficit hyperactivity disorder (ADHD) in children (Mimouni-Bloch et al., 2013). Breastfeeding children is known to have quite positive effects in protecting children against physical illnesses and fostering health and development in many physical and psychological spheres both in the short and long-term. (Current recommendations for breastfeeding are to provide breast milk as the exclusive child's diet for 6 months followed by an additional 6 months that can be supplemented with solid or other foods [American Academy of Pediatrics, 2012].) This study examined whether there is a relation of breast feeding and psychiatric disorder in children.

Three groups were evaluated:

1. A group of children 8–12 years of age diagnosed with ADHD
2. Their siblings who did not show ADHD
3. Children without ADHD who attended an Otolaryngology clinic (ear, nose, and throat problems)

This is an interesting design because group 2 helps to control for common environments of siblings and could help rule out some interpretations other than breast feeding. Group 3 cleverly addresses children who are coming to a medical facility (as were ADHD children) but for a nonpsychiatric reason. Parents were asked to complete measures of breast feeding practices of their children for all groups and to do so for different time periods (e.g., at 1 month after the child's birth, at 3 months, and so on) to get a picture of patterns of breastfeeding. These measures were of course retrospective because the children were now 8–12 years old.

The main finding: lower rates of breastfeeding were associated with higher rates of ADHD. For example, at 3 months after birth, 43% of children with ADHD were being breastfed (Group 1); at this same time approximately 70% for the combined others (groups 2 and 3) were being breastfed. This difference continued at 6 months (29% for ADHD; over 50% for other groups combined). From this study, we can conclude that breastfeeding is indeed related to later onset of ADHD. This is a methodology text, so we should be especially careful in what we say about conclusions. The conclusion is better phrased as follows: ADHD in children is significantly related to parent retrospective report of breastfeeding. This does not change the finding very much, but we did not observe or get current reports of breastfeeding. The study is important in raising the prospect that feeding practices play a role in later onset of ADHD. Subsequent work can now test additional hypotheses about paths, mechanisms, and moderators that might explain or elaborate the connection between breastfeeding and ADHD.

The retrospective nature of the study raises cautions as well, of which the authors were aware. Breastfeeding (A) and ADHD (B) or other variables (C) have relations whose temporal ordering is not resolved by the design. The conceptual view underlying the study is a type breastfeeding precedes ADHD (i.e., A —> B), but from the standpoint of the design, it is possible that the results could be explained another way (i.e., B —> A). Maybe overly active and more difficult children (hyperactivity is also related to less easy temperament, more oppositional behavior) lead parents to not breastfeed or to stop breastfeeding sooner. It is easy to envision not breastfeeding a child who is difficulty to manage or contain or is just a little squirmy. That would explain why their siblings had higher rates of breastfeeding. Also, it is possible that recall of the parents might be biased in a way that supports the finding. Perhaps parents recall their currently ADHD child as more of a problem early in life (whether or not they were) and were more likely to report not breastfeeding them as a result. That is, their recall of breastfeeding may be partially tainted or biased by the children's current diagnosis. In short, the direction of A to B or B to A is not entirely clear.

In addition, other variables not included in the study (C) may well lead to both lower rates of breastfeeding and higher rates of ADHD. I have mentioned temperament—a biological disposition related to how an individual responds. Among the many ways to classify this, easy-to-difficult is one dimension that is meaningful and relates to later psychiatric dysfunction. Infants with a more difficult temperament are fussier, are more likely to resist or cry when handed from one adult to another, more active, a little more intense, and slower to adapt to change. The easy temperament is at the other side of the continuum and is what you (reader) and I obviously were when we were infants. We already know that temperament influences the onset of psychiatric disorder

years later. Perhaps this third variable (C) accounted for both reduced breastfeeding and increased rates of ADHD. Perhaps, there was something else that varied for ADHD children that could explain the findings. Curiously more children in the ADHD group were born prematurely and perhaps that too influenced breastfeeding and later ADHD onset.

The point is that one cannot know the relation of these influences from the design of the study. With all of these interpretations, is there any value to the findings? Absolutely! Knowing that breastfeeding is related to the onset of psychiatric disorder is critical. All sorts of other questions are raised:

- Is it the breastfeeding or some other variable?
- Is there a moderator that is of interest here (e.g., boys vs. girls, single vs. two parent families)?
- What is the mechanism that might be involved?
- Does breastfeeding augment brain development, and also whether ADHD has some characteristic deficits in some areas of the brain?
- Are these connected?

There are intervention issues too. There are international efforts to increase breastfeeding because of the broad benefits. Also, what about prevention? Can research (true-experiments, quasi-experiments with humans and nonhuman animal studies) show that certain deleterious outcomes (ADHD or in animals some of the impulsive characteristics that might be evaluated) are averted with breastfeeding or longer periods of breastfeeding? Again, all questions emanating from the retrospective study are merely a sample of the questions the findings raise.

7.2.3: More Information on Retrospective Design

> Obviously a key issue in retrospective designs pertains to the assessment. As a general rule, retrospective reports permit the investigator to identify correlates. One of the correlates may be recall of a past event which is why I emphasized the conclusions of "parent report of breastfeeding" rather than "breastfeeding."

There can be significant problems that usually preclude establishing the recalled event as a risk factor (antecedent) for the outcome of interest. First, selective recall, inaccurate recall, and recall biased by the outcome (e.g., dysfunction) all interfere with drawing valid conclusions about the past event, its occurrence, or differential occurrence for groups that vary in a later outcome. I mentioned previously that memory is a matter of recoding rather than recording events and experiences that happened previously (Roediger & McDermott, 2000). Thus, recall has some inherent limitations. In some cases, historical records (e.g., school truancy, participation in high school activities) are used as the data. With such records, the quality, reliability, and completeness of the data also raise potential interpretive problems.

All retrospective measures are not necessarily flawed, and hence they ought not to be cast aside categorically when evaluating a study. There are different methods of retrospective assessment (e.g., self-report, archival records), types of events that are assessed (e.g., parenting practices, death of a relative), time frames (e.g., recall of events or experiences within the past week vs. past 25 years), and means of soliciting or prompting the recalled material. These are not all subject to the same sorts or degrees of bias. As a general rule, retrospective reports of psychological states (e.g., family conflict, mental health, difficulties of childhood) and duration, level, and dates of particular events are rather poor; recall of discrete events (e.g., changes in residences) and more stable characteristics (e.g., reading skills) tends to be more reliable but still not great.

For example, longitudinal studies have asked people to evaluate some characteristic at a concurrent point in time; years later retrospective data are obtained on the same characteristic. The results indicate little relationship between the concurrent and retrospective data when the same person evaluates the same events. The implications can be enormous. For example, the prevalence of mental illness (proportion of individuals with a psychiatric disorder up to age 32) was almost two times greater when measured prospectively (at different points over time) than when individuals evaluated their functioning retrospectively (Moffitt et al., 2010). In another study, only 44% of individuals who had met criteria for diagnosis of depression at or before the age of 21 recalled a key symptom, yet the entire disorder, when assessed just a few years later (age 25) (Wells & Horwood, 2004). Similarly, in an example I mentioned previously, the sexual activity of adolescents who took a virginity pledge was compared to those who had not taken the pledge (Rosenbaum, 2009). Five years later 82% had denied taking the pledge.

In some cases, critical information to classify individuals or to evaluate relations among variables relies on recall. Even what might seem to be memorable events are not recalled reliably. For example, a study with girls comprising four cohorts (age groups 11, 13, 15, and 17) was assessed over a 3-year period to evaluate their age of menarche (first menstrual cycle) (Dorn, Sontag-Padilla, Pabst, Tissot, & Susman, 2013). Direct interviews by a clinician (3 times a year for 3 years) and phone calls interviews (quarterly for 9 quarters) by research assistants assessed self-reported age of menarche over time. The in-person clinician interview format yielded more consistent data in recall over the 3-year period. Even so, even evaluations were on average 0.7 years different in recalling onset of menarche for the interviews and 2.2 years for

phone interview data. Here is a good example because different methods were compared, both showed variation in the timing of an event, and one was clearly better than another.

In the previous example, the recall focused on when an event happened. Much work focuses on whether something happened. In a review of retrospective assessments across diverse areas of research sexual abuse, physical abuse, physical/emotional neglect or family discord, false negatives were much more common (i.e., not recalling events that from prior records were known to have happened) than false positives (i.e., recalling something as happening when in fact it did not) (Hardt & Rutter, 2004). Major adversities in one's past are more likely to be recalled accurately.

> Retrospective assessment does not necessarily mean the information is distorted. Both how and what is assessed make a difference.

For example, over time, a small group of children who showed evidence of autism no longer show the signs and are functioning well (Fein et al., 2013). The percentage is not yet known. A case-control study identified and compared three groups:

1. Children with a prior diagnosis of autism but functioning very well (i.e., in regular classrooms, no longer meeting criteria for the disorder).
2. Children with high-end functioning of autism who still met criteria for the disorder (but matched on age, sex, IQ with the prior group).
3. Typically developing peers. Perhaps the diagnoses that groups 1 and 2 had were inaccurate, and group 1 may not even have been accurately diagnosed.

The diagnostic records were reviewed by experts unaware of which groups the children were placed and used to confirm the early status.

The point: Retrospective assessment can mean going back to documented or documentable information that may have other problems (sloppy or incomplete records) but do not have the retrospective assessment recall problem.

In the case of this study, optimally functioning children who formerly were diagnosed as autistic showed no problems with language, face recognition, communication, and social interaction. The optimal functioning group had a history of milder social deficits and the high functioning autism but had been similar in severity of other symptoms (e.g., repetitive behavior, communication deficits). This study is an important breakthrough by beginning to identify that some children with diagnosis will turn out fine, but what percentage, how, why, and when—all to be studied. Overall, what conclusions to draw?

> Retrospective assessment is not inherently flawed but can vary as a function of what is assessed (e.g., major life events) and how it is assessed (e.g., self-report, records).

Comparisons of the same sample assessed at one point and then asked to recall the event years later often reveal not recalling what they had noted years earlier. Major programs of research (e.g., National Comorbidity Study, National Comorbidity Survey Replication Study) have used retrospective assessment effectively in identifying rates, course, and predictors of mental illness, and data from such studies would be prohibitive to obtain prospectively. Among the advantages is that retrospective assessment can be checked in many studies (Kessler et al., 2004). As any assessment retrospective can have significant problems, retrospective assessment can, but does not necessarily, have significant problems and that is important to remember—well maybe not to remember because recall is limited, so perhaps write it down. Self-report is subject to recall biases but again not all retrospective assessment relies on self-report.

7.2.4: Considerations in Using Case-Control Designs

There are strengths and weaknesses of case-control designs (summarized in Table 7.2).

Among the strengths:

1. The designs are well suited to study conditions that are relatively infrequent. In clinical psychology, groups with particular disorders, personality characteristics, or exposure to particular experiences would be difficult or impossible to obtain from sampling a population randomly or from following a community population over time until individuals showed the characteristic of interest. For example, there is keen interest in understanding characteristics of individuals:

   - With a diagnosis of depression, bipolar disorder, schizophrenia, and of course other disorders
   - Who have engaged in abuse of some kind or who have been victimized by it, who are altruistic, heroic, model parents, spouses, gifted, and so on

   > A case-control study identifies individuals with and without the characteristic and asks how the individuals are alike and different from controls and what are some of the reasons they may have reached this particular outcome. This type of research and the research question are excellent both to test theory and to generate theory.

   Once a profile of characteristics is evident that differentiate groups, one can begin to refine the focus and consider why the relations exist, i.e., developing a little theory.

2. The designs are feasible and efficient in terms of costs and resources. The investigator selects the sample and makes the comparisons between cases and controls at a single point in time (i.e., now). The designs do not

involve following samples prospectively, so there is not a long delay in answering questions from the research. Longitudinal research, while methodologically advantageous for all sorts of reasons, as noted in the next section, is costly in time and personnel.

3. The loss of subjects, a constant concern in longitudinal studies, is not a problem in the usual case-control design. Subjects are assessed at a single point in time, usually in one assessment session.
4. Case-control studies can go well beyond merely showing that two (or more) variables are correlated. The magnitude and type of relations (e.g., direct and indirect relations) can be studied and different patterns of relations within a sample can be delineated. Identifying subtypes within a sample for example occurs when the variables of interest correlate differently for one type of case (e.g., males vs. females) rather than another. These differences are considered as moderator variables and can lead to hypotheses about different types of onset and clinical course.

7.2.5: Further Considerations in Using Case-Control Designs

There are weaknesses of the designs as well. They are:

1. The designs demonstrate correlations, and the direction of the relation between one characteristic and another may not be clear at all. Whether those who are happy in their marriages because of unhappy childhood, for example, may be true, but in a case-control study, even if retrospective, it is always possible that the childhood experience is recalled in a special way in adulthood because of marital unhappiness or a feature associated with marital unhappiness (e.g., depression). In a case-control study, there is inherent ambiguity in how the characteristics of interest relate to each other (i.e., which came first, whether they were caused by some other variable). The retrospective study too is usually limited, unless there can be some certainty that the antecedent occurred prior to the outcome. Here too "outcome" is tricky. For example, in the breastfeeding-ADHD study one could say the outcome (second event to occur) was the diagnosis of ADHD between the ages of 8 and 12 and clearly breastfeeding must have come before that. Not really, early signs of ADHD (e.g., impulsivity, high activity) may have preceded early cessation of breastfeeding. Just because the study focused on ages 8–12 does not mean the problem behaviors that comprise the diagnosis did not in some form occur earlier.
2. A causal relation between one characteristic (cognitions) and another (e.g., depression) cannot be demonstrated in case-control designs.

   Even though case-control designs are not well suited to demonstrate causal relations, they are often very good at generating hypotheses about them. The hypotheses can be bolstered by various analyses within the study that help to rule out other influences (e.g., socioeconomic status [SES], physical health) that might be plausible explanations for the finding.

   Also, dose–response relations (showing that the variables are related in a way consistent with a causal hypothesis) can be helpful.
3. There are sampling biases that may influence the relation between the characteristics of interest. Selection of cases and controls may inadvertently draw on samples in which the relation is quite different from the relation in the general population. For example, if one is interested in studying women who are abused by their spouses, one can identify cases at a women's shelter and compare them to a control group (e.g., in the community or from another clinic but who have not been abused). The goal may be to identify whether abused women, compared to controls, have fewer social supports (friends and relatives on whom they can rely). Although the women in the case group may in fact be abused, they may not represent the larger population of abused women who do not go to shelters. Indeed, most abused women do not go to women's shelters; many of these women do not even consider themselves to be victims of abuse. In addition, the absence of a support system (and other characteristics such as the level of stress) may influence who comes to shelters so that this is a unique group. That is, the lack of social support may actually relate to who comes to shelters to begin with. Consequently, the correlation between abuse and social support may be spurious because support influenced the referral process, i.e., who comes to a shelter. Stated more generally, how cases are identified can greatly influence the relations that are demonstrated within the data. If a special sample is identified because they have self-selected by volunteering to come to a clinic facility or have been directed to do so (e.g., court ordered), the relations that are demonstrated may have little generality to the larger population of interest. It is for this reason that epidemiological research, where these designs are commonly used, relies heavily on random sampling from the general population to identify cases and controls.
4. On balance, the design strategy is extremely valuable. Apart from elaborating concurrent and past characteristics associated with a given problem, characteristic, or facet of functioning, the designs can identify relations among multiple influences. Related, among

multiple variables that might be studied, the magnitude of the relations and variation in the relations as a function of other variables such as sex, age, or race may be very important.

5. The designs often permit investigation of phenomena not easily studied in experiments—we cannot expose individuals to experiences on a random basis or "give" people diagnoses randomly.

Table 7.2: Major Strengths and Weaknesses of Case-Control Designs

Strengths	Weaknesses
Well suited to studying conditions or characteristics that are relatively infrequent in the population	No time line is shown among the variables of interest (e.g., depressed [A] individuals have a certain type of cognitive style [B]), so one cannot usually establish whether one characteristic (A or B) preceded the other or emerged together
Efficient in terms of resources and time because of the cross-sectional assessment	Causal relations cannot be directly demonstrated, even though various analyses (e.g., dose–response relations) can provide a strong basis for hypotheses about these relations
No attrition because of assessment at one point in time	Sampling biases are possible depending on how the cases (e.g., depressed clients) were identified and whether some special additional characteristic (e.g., coming to a clinic) was required
Can study magnitude type of relations among variables (e.g., direct influence, moderating influence)	
Allows the investigator to match (equalize) subjects on one of the variables assessed at pretest (e.g., level of anxiety) that may influence the results	
Can rule out or make implausible the role of influences that might be confounded with the characteristic of interest	
Can generate hypotheses about causal relations or sequence of characteristics and how they unfold to produce a problem	

7.3: Cohort Designs

7.3 Compare case-control designs with cohort designs

Cohort designs refer to strategies in which the investigator studies an intact group or groups over time, i.e., prospectively.

Cohort is a group of people who share a particular characteristic such as being born during a defined period of time.

In a cohort design, a group is followed over time. The design is also referred to as a prospective, longitudinal study. Two key differences help distinguish case-control designs, discussed previously, and cohort designs to which we now turn are:

- Cohort designs follow samples over time to identify factors leading to (antedating) an outcome of interest.
- The group is assessed before the outcome (e.g., depression) has occurred. In contrast, in case-control designs, the groups (cases and controls) are selected based on an outcome that has already occurred.

The special strength of cohort designs lies in establishing the relations between antecedent events and outcomes. Because cases are followed over time, one can be assured of the time line between events, i.e., that the antecedent occurred before the outcome of interest.

The time frame of a prospective study may be a matter of weeks, months, or years, depending on the goals of the study. In such a study, the antecedent condition is assessed (e.g., birth defects, early attachment, sibling relations), and one is assured that the outcome has not occurred (e.g., school competence, anxiety disorder). That is, the temporal order of antecedent and outcome is clear. Hence, a necessary condition for demonstrating a causal relation is met within the design. Of course, demonstrating that an antecedent condition preceded an outcome, by itself, does not establish a causal relation but provides a critical prerequisite. There are many variations of the design, three are considered here.

7.3.1: Single-Group Cohort Design

A single-group, cohort design identifies subjects who meet a particular criterion (e.g., children exposed to domestic violence, individuals released from prison, youth enrolled in preschool day care for at least 2 years) and follows them over time.

The group is selected to examine the emergence of a later outcome that might be of interest (e.g., physical or mental health problems, alcohol use in adolescents, high levels of achievement in adulthood).

The basic requirements of a single-group cohort include assessment at least at two different points in time and a substantial sample that, during that span of time, changes status on the outcome of interest.

For example, all cases referred to a clinic may be identified and assessed. They are then followed prospectively (e.g., over the next 3 years) to identify who shows a recovery (minimal symptoms, functioning well in everyday life) and who does not. Similarly, all children or a large sample of children who were exposed to a local tragedy (e.g., school shooting in a community) might be followed (e.g.,

over the next 12 months) to identify who experiences symptoms of posttraumatic stress disorder (PTSD). Although the subjects in each example were identified and selected as a single group, following cases over time has as its goal identification of those who have different outcomes, i.e., delineation of subgroups at the point of outcome assessment.

Cohort studies begin with a group and evaluate the group over time. Time frames can vary markedly from months to decades, but most fall within a time frame of 1 or 2 years. Yet for many cohort studies, there is no "one study" with a time frame but rather an ongoing series of studies that draw on different time frames and different facets of the sample that have been studied. Here is a well-known example in clinical psychology that began in a distant past then but continues with extended outcomes that are currently being examined.

Decades ago, a cohort design began to study the impact of a hurricane on children (Hurricane Andrew in Florida in 1992; La Greca, Silverman, Vernberg, & Prinstein, 1996). This hurricane was one of the worst natural disasters in the United States, leaving 175,000 families homeless and without adequate food or supplies and exceeding costs of any other natural disaster at that time (over $15.5 billion). The investigators examined the extent to which the hurricane led to persistent symptoms of PTSD over the ensuing months. These symptoms include:

- Re-experiencing the disaster (intrusive thoughts and dreams)
- Difficulty sleeping and concentrating
- Detachment and avoidance of disaster-related activities

In current psychiatric classification, these symptoms characterize PTSD and reflect impairment that results directly from the experience of trauma or disaster (e.g., exposure to war, rape, or other extremely stressful event).

School children (3–5th grade, N = 442) exposed to the hurricane were identified and assessed over time on three occasions: 3, 7, and 10 months after the hurricane (La Greca et al., 1996). Among the goals was to predict which children showed PTSD symptoms at the final assessment and what factors predicted this from the earlier assessments. The results indicated that PTSD symptoms decreased for the sample over time. At the final (10-month) assessment, 12% of the children continued to show severe symptom levels. The most salient predictors of who showed severe PTSD symptoms were the extent to which the initial disaster was perceived by the youths to be life-threatening and the severity of loss and disruption during and after the disaster (e.g., loss of property, disruption of housing, routines). Greater threat and disruption were associated with more severe PTSD symptoms. Less social support from family and friends, the occurrence of other life events, and high efforts to cope with the trauma (e.g., blame and anger) also predicted persistence of symptoms. These results help understand factors that are associated with persistence of symptoms of trauma and also provide clues of what might be addressed to intervene early (e.g., stabilize the disruption as soon as possible and help restore normal routines) among youths at greatest risk. Of course, we do not know that intervening on the associated features will change the experience of trauma, but the findings provide potential leads.

Although the study began in the 1990s, the data have generated several studies and additional findings. For example, more recent studies of the sample have looked at children with comorbidity (presence of multiple symptoms) and their outcomes, evaluated different trajectories or paths leading to different outcomes, predictors of resilience or not having the untoward outcomes, and among other foci (e.g., La Greca, Silverman, Lai, & Jaccard, 2010; Lai, La Greca, Auslander, & Short, 2012). For these recent studies, the longest follow-up of the children was 21 months after hurricane. Yet, new questions that can be asked of the data set and sample and variations of the outcomes can be evaluated. Overall, the study of the impact of the hurricane nicely illustrates a cohort design and its key characteristics, namely, identifying a group, following the group over time, delineating different outcomes (e.g., remission vs. continuation of symptoms), and identifying antecedent factors that are associated with varied outcomes.

7.3.2: Birth-Cohort Design

A special variation worth delineating is referred to as *birth-cohort design*.

As the name suggests, this is a study that begins with a group that enters the study at birth. There usually is a specific time frame (e.g., 6- or 12-month period) and geographical locale (country, state or province, city, district, hospital). Children born in the specific time period and geographical setting are now the participants. They are then followed for an extended period through childhood and adulthood spanning decades.

Birth-cohort studies often identify multiple domains over the course of life and provide a rich data set for evaluating functioning and precursors of both adaptive and maladaptive functioning.

Sometimes the outcomes of interest are focused (e.g., diagnosis of schizophrenia) but often broad domains are assessed both early in the design and throughout to assess mental and physical health and functioning in multiple domains (e.g., school, work, social relations, society). Assessments are obtained regularly, but the intervals need not be the fixed or the same (e.g., every 12 months) in a given study.

Some of the measures vary at different points as the participants enter into different facets of their life over the course of development. School functioning (e.g., elementary school, middle school), criminal behavior (convictions as a teen or young adult), and unemployment and marital status (e.g., in adulthood) convey some of the obvious domains that are likely to be assessed at different developmental periods.

There have been several birth-cohort studies and often their beginnings are in some distant past (e.g., Esser, Schmidt, & Woerner, 1990; Farrington, 1991; Silva, 1990; Werner & Smith, 1982). Yet, their yield usually continues in the decades that follow with a new set of investigators who take over evaluation of the data set.

For example, a birth-cohort study has been ongoing for some time in New Zealand to understand the development of psychopathology and adjustment (Silva, 1990). The study began by sampling all children who could be identified (N = 1,037) who were born in the city of Dunedin (pronounced "done-EE-din" or if you just finished lunch, "done-eatin") (approximate population of 120,000) within a 1-year period (1972–1973). From the ages of 3 to 15, youth were assessed every 2 years, and then again at ages 18, 21, 26, and in recent reports age 32 (e.g., Moffitt et al., 2010). At each assessment period, participants came to the research setting (within 2 months of their birthday) and completed a full day of assessments (physical exam, mental health interview, and so on) with measures, of course, changing with age of the subjects. Many findings have emanated from this project.

As a recent example, the relation of self-control early in life was evaluated as a predictor of outcomes decades later (when the sample was age 32) (Moffitt et al., 2011). Self-control is an umbrella term that encompasses the ability to delay gratification, control impulses, and regulate emotional expression. In this study, self-control was assessed at different periods (ages 3, 5, 7, 7, 11) and using reports from observers, teachers, parents, and children. Composite measures were made from these indices. Hypotheses were tested related to the relation of self-control to multiple other domains in teenage years and early adulthood and in relation to mediation of these relations.

The main findings, only lightly sampled here, were that:

- Lower self-control measured early in life was reflected in poorer physical and mental health
- Poorer financial status
- Higher rates of criminal activity in young adulthood (age 32)

These outcome measures were composites based on multiple indices:

1. Health included cardiovascular, respiratory, dental and sexual health and inflammatory status and included laboratory tests and results from a physical exam.
2. Mental health was assessed with clinical interviews to evaluate depression and substance dependence (e.g., tobacco, alcohol, cannabis).
3. Financial status included a range of measures such as indices of financial planning and holdings (e.g., owing a home, having a retirement plan) but also struggling as evident with credit problems.
4. Criminal activity included convictions for assorted crimes.
5. Gradients (dose–response relations) were found so that lower levels of self-control early in life were associated with increasingly more deleterious outcomes.

Overall, the study shows that self-control early in life predicted outcomes in multiple domains. Many important questions come to mind for future studies. Perhaps the most salient and raised by the authors is whether intervening to change self-control would have impact on those outcomes. Interestingly, over the course of the study, a small portion of individuals changed in their self-control. They moved in the direction from lower to higher self-control. The reason for the change is not clear. These individuals had better outcomes to whom they were similar before (low self-control levels) but who did not change. This is intriguing but we have to remain methodologically clear. The study shows correlations, as the authors carefully note. Low self-control was a risk factor for poor outcomes. Will it be causally involved, and can it be changed to alter outcomes? These are very important questions for basic research (e.g., malleability of self-control, plasticity of the brain in those regions that govern self-control, animal models of impulsivity and limited inhibition and whether that can be turned on and off in the brain) and of course applied research (e.g., improving health and reducing criminal behavior).

7.3.3: More Information on Birth-Cohort Design

There are few special features to note about birth-cohort studies. These are:

1. The strength of the study derives from comprehensive assessments repeatedly over extended periods. Participants are called back, and usually multiple assessments are obtained using many assessment modalities (e.g., reports from different informants, laboratory tasks, laboratory tests to sample indices of health, community records for measure criminal activity or employment). The repeated assessments and the extensive assessments place a potential strain and burden on the subjects. From the standpoint of the investigation, a large percentage of the sample must be retained; otherwise, selection-biases and poor external validity might result. That means that

investigators and their team must keep in very close contact with the families whether or not an assessment is scheduled for that year.

2. The dataset is without peer in understanding how development unfolds, in identifying the multiple paths toward adaptive and maladaptive functioning. I have covered the obvious by emphasizing assessment of children through adulthood, but "Wait, there's more!" We have learned that one's grandparents (e.g., diet, age when they had their children) affect the health of the grandchild. Birth-cohort studies occasionally evaluate three generations. These generations include:

   - The babies who become children and are followed
   - Their parents who complete assessments over the course of the study
   - The offspring of the babies who are now all grown up. That is, the original birth-cohort grows up, gets married, and has children

   So now the investigators can study the original babies who are growing up, characteristic of their parents and families, and then start much of this again with the grandchildren (e.g., Högberg, Lundholm, Cnattingius, Öberg, & Iliadou, 2013). More generally, research in physical and mental health and certainly other domains (e.g., education, employment) has focused on parent and child (intergenerational) connections. It is clear that multigenerational influences exert their own influences. Birth-cohort often can get at these in a novel way because of the rich assessments over an extended period (see Power, Kuh, & Morton, 2013).

3. Effort, cost, and obstacles (e.g., retaining investigators, cases, and grant support) make birth-cohort studies relatively rare. Obtaining grant support for 30 years or even 5 years is not guaranteed, so there is the problem of keeping funding consistent as government priorities and financial conditions change. Yet, at the beginning of a study, usually there is significant background work and part of that is to have stable funding agreed on in advance. Although such studies are relatively rare, occasionally there are multiple birth-cohort studies on a question. For example, understanding the development of schizophrenia has been of enormous concern, and 11 birth-cohort studies from 7 different countries have been provided to elaborate the paths leading to the disorder (Welham, Isohanni, Jones, & McGrath, 2009).

4. "New" researchers usually are needed. The participants (infants) are likely to outlive the careers of the investigators who started the project. Consequently, a young new investigative team must be woven into the project and take over the data collection, database, publication of the studies, and so on. This means also that as a researcher you may have the opportunity to work on a birth-cohort study that may have started before you were born. In addition, databases for birth-cohort studies occasionally are made available to one who can do research directly from them.

From the standpoint of this chapter, it is important to leave birth-cohort studies. Cohort studies do not necessarily mean *birth*-cohort studies.

> The defining advantage of the cohort study is being able to identify the time line between antecedents and outcomes, and 1 to a few years is the usual time frame for such studies within psychological research. Yet, if you see a study that used a birth-cohort design, chances are that there are scores of studies generated from that same database.

7.3.4: Multigroup Cohort Design

> The *multigroup cohort design* is a prospective study in which two (or more) groups are identified at the initial assessment (Time 1) and followed over time to examine outcomes of interest.

One group is identified because they have an experience, condition, or characteristic of interest; the other group is identified who does not. So far, this description is exactly like a case-control design. A case-control design and two-group cohort designs are distinguished in the following way. A case-control design selects two groups—one of which shows the *outcome* of interest (e.g., is depressed) and the other group which does not (e.g., not depressed).

> A two-cohort design begins by selecting two groups that vary in exposure to some condition of interest or risk factor (e.g., soldiers returning from combat) or not (e.g., soldiers returning from the same locale but who did not experience combat) and follows them to see what the outcomes will be.

As noted before, the distinguishing feature of a cohort design is that cases are *followed prospectively* to see what happens, i.e., the outcomes that emerge.

A classic example from developmental psychopathology is worth retrieving from the past because of its continued relevance but also the exemplary methodological thinking behind the design and assessments. This two-cohort design was used to determine whether a head injury in childhood increases the chances of later psychiatric disorder (Rutter, 1981; Rutter, Chadwick, & Shaffer, 1983). The hypothesis was that brain damage is one factor that can lead to psychiatric disorders later. Youths who received head injury (e.g., accident) were identified and assessed over time for a 2-year period. The obvious control group would be a sample of youths without a head injury, matched on various subject (sex, age, ethnicity) and demographic variables (e.g., social class) that are

known to influence patterns of psychiatric disorders. However, a noninjury group may not provide the best comparison or test of the hypothesis. The hypothesis focused on *head* injury. Maybe the experience of *any* injury would increase later psychiatric impairment. Perhaps any injury (whether to the head or toes) that leads to hospitalization for a child (or anyone) is traumatic and that trauma and entry into a hospital alone could increase later impairment. Even if a head injury group showed greater subsequent psychiatric impairment, that would not be a strong test of the hypothesis. There would be a construct validity problem—injury or head injury? In this study, the second group (making it a two- or multi-group cohort design) consisted of youths who were hospitalized for orthopedic injury (e.g., broken bones from accidents). Thus, both groups experienced injury, but head injury was the unique feature of the index group expected to predict later psychiatric disorder. Both groups were followed for 2 years after the injury and evaluated at that time.

As predicted, the results indicated that youths with head injury had a much higher rate of psychiatric disorder at the follow-up 2 years later when compared with orthopedic injury youths. The study might end here and is still considered to support the original hypothesis. However, more was accomplished to strengthen the inferences (construct validity) that could be drawn:

1. One interpretation of the results is that children who get head injuries are not a random sample of youths in the population. Perhaps they already have more psychological and psychiatric problems to begin with (i.e., before the head injury). In fact, emotional and behavioral problems among children are correlated with more risky and impulsive behavior, which could increase the risk of head injury. Showing that a head injury group, when compared with another group, has higher a rate of psychiatric disorder would not establish the temporal order of head injury and later psychiatric disorder. The goal of this study was not only to show that injury was related to later psychiatric impairment but also to establish that it preceded such impairment. Collection of retrospective data during the study helped address this. Immediately after the injury, families of both head and orthopedic injury group children completed assessments that evaluated pre-injury emotional and behavioral problems of the children in both groups. Pre-injury problems did not differ between groups nor predict later child psychiatric impairment. Thus, it is unlikely that preexisting psychological problems could explain the relation of head injury and later psychiatric disorder.
2. If brain damage were the key factor, one hypothesis would be that severity of the injury and subsequent incidence of psychiatric disorder would be related. As mentioned previously, observational studies often look for a dose–response relation within the index or case group to see if there is a gradient in the association between the amount of one variable and the rate of the outcome. The presence of a dose–response relation is one more bit of evidence suggesting that the construct of interest is the key in explaining the outcome. In this study, severity of brain injury was considered to provide a further test of the hypothesis. As a measure of severity of brain injury, the authors used the number of days of postinjury amnesia (not remembering the incident). Youths with more days of amnesia (≥8 days), compared with those of few days of amnesia (≤7 days), showed much higher rates of later psychiatric impairment. This further suggests that the construct, head injury, is likely to explain the relation.

7.3.5: More Information on Multigroup Cohort Design

Overall, noteworthy features of this study are the use of a comparison group that helped evaluate the specific role of head injury, the use of assessment (albeit retrospective) to address one threat to construct validity (that group differences were due to preinjury emotional and behavioral problems), and data analyses (dose–response relation) to suggest further that head injury was the likely variable accounting for the follow-up results.

Does this study establish that head injury is a *cause* of psychiatric disorder?

What do you think?

The study did establish that head injury preceded psychiatric disorder and hence one condition of a causal relation was established, namely, a time line where the proposed event (cause) comes before the outcome. Further analyses also establish that head injury (rather than just injury) was the likely influence. At the same time, we cannot be absolutely certain that there is a causal relation. It could be that some other construct not assessed in this study is the factor and head injury is not main variable.

For example, children vary in the extent to which they are clumsy early in life, as defined by motor movement and coordination. Clumsiness in early childhood is a predictor of later psychiatric impairment as known from several studies (Fryers & Brugha, 2013). It is possible and plausible that the head injury group varied (from the orthopedic group) on clumsiness. Perhaps, head injury was merely a correlate of this clumsiness and clumsiness is the key factor. No doubt we could generate other explanations, all a matter of surmise and

further research. The study cannot rule out all other causes. Yet, the careful selection of controls, assessment, and data analyses act in concert to reduce the plausibility that other factors than head injury were responsible for the findings. The original study went very far to establish that head injury plays a role. Additional research might unravel whether the effect is a direct influence (i.e., injury harms brain functioning that disrupts social, emotional, and behavioral processes) and/or indirect influence (i.e., head injury leads to other processes, perhaps in the family, leading to disorder). The study stands as an excellent example of a multi-cohort design as well as a model of methodological thinking. That thinking is based on the investigators considering what they wanted to talk about (head injury), what might be rival explanations, and what could they do to make some of those rival explanations less plausible than what they wanted to talk about.

7.3.6: Accelerated, Multi-Cohort Longitudinal Design

An accelerated, multi-cohort longitudinal design is a prospective, longitudinal study in which multiple groups (two or more cohorts) are studied in a special way.

The key feature of the design is the inclusion of cohorts who vary in age when they enter the study.

The design is referred to as accelerated because the period of interest (e.g., development of children and adolescents over the course of 10 years) is studied in a way that requires less time (e.g., less than 10 years) than if a single group were followed over time.

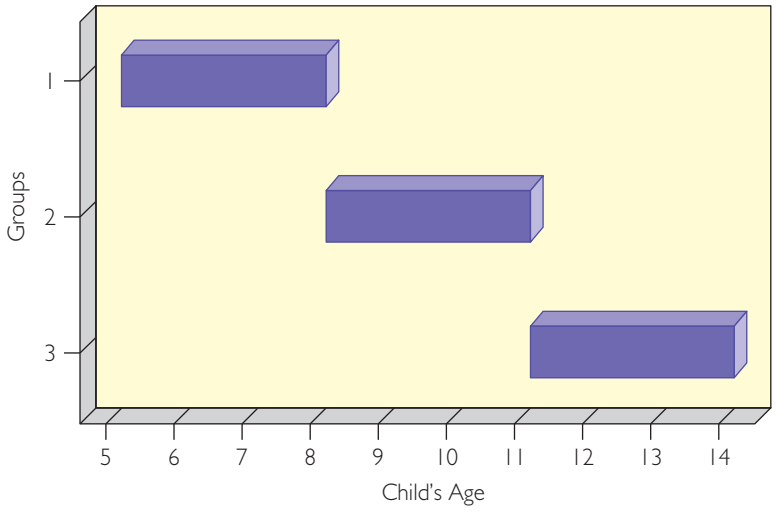
This is accomplished by including several groups, each of which covers only a portion of the total time frame of interest. The groups overlap in ways that permit the investigator to discuss the entire development period.

Consider an example to convey how this is accomplished. Suppose one were interested in studying how patterns of cognitions, emotions, and behavior emerge over the course of childhood, say from ages 5 to 14, a period that might be of keen interest in light of school entry, school transitions, and entry into adolescence. Each of those periods has its own challenges from early socialization to risky behaviors in adolescence. An obvious study would be to identify one group (a cohort) and to follow them from first assessment (age 5) until the final assessment when they become 14. That would be a single-group cohort design, as discussed previously. Another way would be to study the question with an accelerated, multi-cohort longitudinal design. The study could begin with three groups that vary in age. For this example, let us say that the three groups we identify are ages 5, 8, and 11 years old. Each group is assessed at the point of entry (when we start) and then followed and assessed for the next 3 years. Assume that assessments are conducted annually during the month of each child's birthday.

Figure 7.1 diagrams the study with three groups to show that each group is assessed for a period of 4 years beginning at the point of entering the study.

Figure 7.1: Accelerated Multi-Cohort Longitudinal Design

An accelerated, multi-cohort longitudinal design in which separate groups are selected and assessed. Their ages span the entire period time frame of interest (ages 5–14) but no one group is followed for the entire duration. Time 1 (first assessment) is when the youths are 5, 8, and 11 years of age, respectively.



> There is a cross-sectional component of this design that consists of comparing all youths at the time they first enter the study and are at different ages.

Also, we are interested in comparing the 5-year-old group when they become 8 years old with the data from the 8-year-olds when they entered the study to see if the two groups are similar on the measures. That is, there are two 8-year-old groups at some point in the design and one can see if the data are similar from different cohorts when they are the same age.

> The longitudinal component of the design examines development over the period of 5–14 years of age. By seeing how each cohort develops and the relations over time within a group, one hopes to be able to chart development across the entire period from ages 5 through 14, even though no one group was studied for the entire duration.

The example conveys only one way of selecting groups. The number of groups, the assessment intervals, and the overlap among the groups over the course of development can all vary.

7.3.7: More Information on Accelerated, Multi-Cohort Longitudinal Design

There are two salient issues that an accelerated longitudinal design is intended to address. **First**, the design can identify if the characteristics of a particular cohort are due to historical influences or special features of the period in history in which the cohort is assessed. Consider this potential artifact. In a single-group cohort design, a group is followed over an extended period. It is quite possible that the information generated by the group is special in light of the period in time in which the study was completed. For example, if one is interested in studying the relation of factors that occur during the course of adolescence to outcomes in young adulthood, obviously a longitudinal design can begin by identifying adolescents and assessing them repeatedly at various intervals until they become adults.

The data may reveal patterns among the measures (e.g., correlations among key characteristics), changes over time, and factors that predict particular outcomes that are unique. There is a possibility that the results might be attributable in part to the period in which the individuals have been studied; that is, this cohort may show special results because of being youths who grew up during a period with or without the availability of some factors that might influence the variables that are studied. Influences that could affect a given cohort and many output of interest (e.g., violence, marital happiness of that cohort) are:

- Changes in the availability of smart phones
- Easier to use methods of birth control
- The availability of two parents in the home (low rate of divorce)
- Unemployment rates in the country (which affect individual families)

Two examples of cohort effects are the prevalence of tattoos and use of marijuana (medicinal of course). The prevalence of both of these was relatively low a few decades ago but is much more common now and mainstream in many circles.

Characterizing individuals at one point in time (e.g., those long decades ago) would readily be expected to yield differences from those who had tattoos and who consumed marijuana.

> The term "cohort effect" refers to characteristics that are associated with different groups and different periods of time. People in everyday life understand cohort effects. Grandparents and parents (and eventually you, the reader) invariably begin sentences with phrases like, "When I was a child . . . ." or "When I was in college . . . ."

This sentence gets filled in with some practice (e.g., taking a stage coach to school, showing obsequious respect for an elder person, not thinking of undressing in front of romantic partner until 5 years into marriage). Any sentence beginning that way means the person is referring to a cohort effect, i.e., things were different then.

More generally, culture and its practices and values are always changing (e.g., unemployment and crime rates, wars, values), and these historical events can influence the pattern more than any particular cohort shows. Thus, in a single-group cohort design, it is possible that the group shows a pattern that is influenced in critical ways by events occurring during this period (i.e., history as a threat to external validity). The results (relations among variables, developmental paths) may differ if another cohort were studied at a different period or point in time.

> An accelerated, multi-cohort design allows one to better separate any historical period effects from developmental change. Each cohort within the study has a slightly different history and one can make comparisons to address whether there are period influences.

In the example (Figure 7.1), the investigator can compare the data of the 5-year-olds when they turn 8 years of age with the data of 8-year-olds. These groups ought to provide similar information, namely, how 8-year-olds are on the measures of interest. Major differences at this point raise the prospect of some other broad historical influence that is at work. In any case, one advantage of an accelerated longitudinal design is the

ability to evaluate whether the findings for the cohort are restricted to possible historical influences that are unique to that group.

Second and more obvious, the accelerated longitudinal design addresses the most difficult part of longitudinal designs, namely, they take an extended period to complete. The period can be reduced by using multiple cohorts to represent different and overlapping periods of that time frame. In the example in Figure 7.1, the goal was to study development covering a period of 10 years. Using an accelerated design, each of the three groups in the example was assessed over a 4-year period, although the 10 years of interest was examined. In making the study shorter, some of the problems of longitudinal research (attrition, expense of following and finding cases) are likely to be reduced.

7.3.8: Considerations in Using Cohort Designs

Cohort designs have their strengths and weaknesses, and these are highlighted in Table 7.3.

As to the strengths:

1. The time line between proposed antecedents (risk factors, causes) and the outcome of interest can be firmly established. This is not a minor point and serves as the primary basis for distinguishing the variations of observational designs (case-control vs. cohort designs) we have discussed.
2. Careful assessments can be made of the independent variables (antecedents, predictors) of interest. Because the outcome of interest has not yet occurred, one can be assured that the outcome did not bias the measures. Measurements at Time 1 (and other occasions) will not be influenced by the outcome, which will not be determined until much later at Time 2.
3. Because the designs are prospective and assessments are made on multiple occasions, the investigator can plan and administer measures that will thoroughly assess the predictors (e.g., multiple measures, multiple methods of assessment) at the different points in time. A given influence may be assessed on more than one occasion and the accumulation of different influences over time can be examined as predictors of an outcome.
4. Cohort designs are good for testing theories about risk, protective, and causal factors. My comments have focused on merely describing relations and that is critical. But one can test theory, make predictions, and offer explanations of what is and is not involved in a particular outcome and how multiple variables may combine. These can be tested in cohort designs.

Table 7.3: Major Strengths and Weaknesses of Cohort Designs

Strengths	Weaknesses
Can firmly establish the time line (antecedent becomes before some outcome of interest)	Prospective studies can take considerable time to complete, and answers to critical questions (e.g., effect of asbestos and smoking on health, effect of physical or emotional abuse on youths) may have delayed answers
Measurement of the antecedents could not be biased by the outcome (e.g., being depressed now could not influence past recall of events early in life—those events were assessed before being depressed)	Studies conducted over time can be costly in terms of personnel and resources. Retaining cases in a longitudinal study often requires full-time efforts of researchers in the study
Multiple methods and assessments at different points in time can be used to assess the predictors to chart the course or progression from the antecedent to the outcome	Attrition or loss of subjects over time can bias the sample
All of the permutations can be studied in relation to the antecedent (occurred or did not occur at Time 1) and outcome (subjects did show or did not show the outcome at Time 2)	Cohort effects may serve as a moderator, i.e., it is possible that the findings are due to the sample assessed at a particular point in time
Good for generating and testing theory about risk, protective, and causal factors and mediators and moderators	The outcome of interest (who becomes depressed, engages in later criminal behavior, and commits suicide) may have a relatively low base rate. Statistical power and sample sizes become issues to evaluate the outcome

7.4: Prediction, Classification, and Selection

7.4 Analyze how prediction, classification, and selection are ways of referring to some outcome

Another strength of cohort designs, and observational designs more generally, pertains to the interest in varied outcomes for different groups as well as prediction, classification, and selection of cases.

7.4.1: Identifying Varying Outcomes: Risk and Protective Factors

Different emphases of this strength in diverse outcomes can be delineated. First, consider a prospective longitudinal two-group design. We select two groups to begin a study. One group has had an experience of interest to us and another group has had no exposure. Among the many strengths of a prospective, longitudinal study is the ability to examine the full set of possibilities among those who do and do not experience antecedent condition and those who do and do not show the outcome.

For example, consider the hypothesis that watching videos high in aggressive behavior in early childhood is associated with later aggressive behavior in adolescence. Assume for a moment that we will conduct this study with a two-group cohort design and we have selected 500 children in a community aged 6–8 years. We follow these children for 10 years and evaluate their aggressive behavior (fighting at school). For simplicity sake, let us classify exposure to video aggression and later aggressive behavior in a dichotomous fashion, even though we know that each of these is a matter of degree (dimensional). So let us say, at Time 1 (childhood) we can identify children who are exposed to high levels of videos with aggressive behavior (e.g., killing, decapitating, and destroying others) or not exposed to aggressive videos at all (two groups). This makes the study a two-group, cohort design. At Time 2 (adolescence), let us identify the outcome as high in aggression at school or not (two outcomes).

We can divide the cohort into two subgroups based on these combinations. The subgroups (Cells A, B, C, and D) are diagramed in Figure 7.2 and described here:

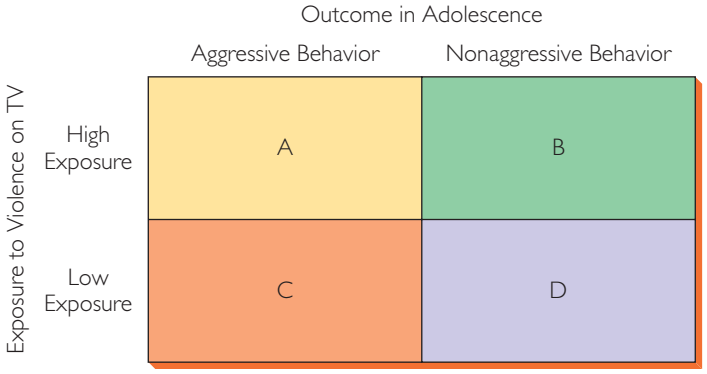
- Those who *experienced the antecedent* in childhood (exposed to high levels of TV aggression) and *the outcome* (they are high in aggression in adolescence).
- Those who *experienced the antecedent* (exposed to high levels of TV exposure), but *did not show the outcome.*
- Those who *did not experience the antecedent*, but *did show the outcome.*
- Those who *did not experience the antecedent* and *did not show the outcome.*

Based on this initial assessment, youths are classified as exposed to aggressive television or not exposed to aggressive television. They are then followed prospectively. Typically in such research, assessment continues on multiple occasions (e.g., every year or few years), but in this example we are considering only time 2 assessment at some later point in adolescence. In adolescence we assess all cases and classify them at that point on whether they are exhibiting aggressive behavior. The four groups resulting from the design are delineated in the cells.

The four cells in Figure 7.2 convey one of the strengths of a prospective design. The design allows one to evaluate whether exposure to video aggression in fact has higher rates of later aggression but has many other interesting possibilities. For example, in Cells A and B, we have all of the children exposed to aggressive videos. Some of these children became aggressive later (Cell A) but others did not (Cell B). Comparing these individuals on a host of antecedent conditions may *suggest* why individuals who are exposed do not develop aggression later. The comparison conveys the correlates of these different outcomes (e.g., individuals who did not show aggression as expected were more x or y when they were younger). This can be very useful in generating hypotheses about why individuals did not become aggressive in adolescence. Also, we can look at those children who were not exposed to aggressive videos at all. Some of these children became aggressive anyway (Cell C) but others did not (Cell D). What factors are involved in developing aggression in adolescence among youth who have not been exposed to video aggression? Measures obtained before the outcome that are available in the study may shed light on these questions. I have not elaborated all of the comparisons of interest. Yet, the larger point can be made, namely, that an advantage of a prospective study is evaluation of the rates of the onset of some outcome in the cohort of interest and exploration of factors that increase or decrease the likelihood of the outcome, based on comparisons of subgroups who vary on the presence (or degree) of the antecedent condition and the presence (or degree) of the outcome.

The questions embedded in the four cells I have illustrated are often intriguing and suggest further lines of

Figure 7.2: Hypothetical Two-Group Cohort Design



research. For example, how old was your mom's father (i.e., your grandfather on your mom's side) when your mom was born? Grandfathers who become parents to daughters when they are 50 years of age or older have grandchildren who are much greater risk for autism than grandfathers who became parents when they were in their 20s (Frans et al., 2013). But let us look at our four cells again. We have two levels of grandfathers (have their children when under vs. over 50) and two outcomes (later grandchildren who were diagnosed with autism spectrum disorder [ASD] and those who were not). Among the many questions, for grandfathers who were over 50 when their daughters were born, some did (Cell A) and some did not (Cell B) have grandchildren with ASD. It would take a while and many studies to work that out, but in the process we could elaborate additional influences that increase or decrease the likelihood of the outcome. Not included in the design of the study or my comments is more information about the findings. We do not merely identify associations (e.g., whether grandfather age does or does not increase risk) but the magnitude of the relation. In the autism example, grandfathers when over 50 when their daughters were born had a 1.67 greater chance of having a grandchild with autism compared with the grandfathers who had their children when younger. That is over 1½ times greater risk but still does not tell us how many out of 100 grandchildren we would expect to show later ASD.

In psychology, considerable research has been done using longitudinal designs with the four Cells illustrated in Figure 7.1. For example, consider comparing two groups (Cells A and B again). Individuals in both cells had the experience but those in Cell B did not show the problem. What made that happen? Is there some other variable they have that might explain why they did not get the problem? That other variable is referred to as a *protective factor*, a concept discussed previously. For example, youth who are at high risk for delinquency but who do not become delinquents often have a significant adult (e.g., coach) in their lives to whom they relate well and that serves as a protective factor. It would be a complete methodological nonsequitur to think that giving at-risk delinquents someone to relate to would decrease their delinquency.

> A protective factor is a correlate and could be a proxy (stand for) for some other variable. For example, perhaps children who have a positive relation with an adult may be less obnoxious in general and could form such relationships—the protective feature is not in the other adult relation but in the child's attributes. Such explanations can be addressed in future research.

Even so, it is useful to identify protective factors. Some of those may be malleable through psychological intervention and some of those might in fact bear a causal relation and protect individuals. These are critical questions in clinical psychology and are addressed in observational designs, especially cohort designs.

7.4.2: Sensitivity and Specificity: Classification, Selection, and Diagnosis

We have discussed cohort designs in which there is interest in evaluating the onset or occurrence of a particular outcome. More broadly, research is interested in classification, selection, and diagnosis—all ways of referring to some outcome in which we are interested. Prediction and selection of outcomes are fundamental to clinical psychology but to so many other disciplines as well (e.g., public health, medicine, criminology, homeland security, business, and advertising). We use research to identify variables that predict an outcome and all sorts of variables (e.g., genetics, early experience, diet, and so on). Among the goals is to identify or classify those individuals who show a particular outcome or engage in some behavior at a later time. This was covered in the prior discussion of risk and protective factors, but this discussion has a slightly different thrust.

As researchers but also as citizens, we are deeply interested in classification. For example, at airports, security agents are interested in identifying terrorists—that is a classification challenge—look at everyone and pluck (classify) those who are likely to terrorize. In national governments, federal tax agencies are interested in identifying who is most likely to cheat on one's tax reports and those individuals are more likely to be scrutinized (auditing of people's tax reports). The variables used to make the classification for terrorists or tax evaders and the precise predictive weight each variable is given are secrets, but we can guess likely candidates. But beyond the secretive questions, there are many more instances in which we want to use observational data (e.g., screening, assessment) to classify individuals into groups. For example:

- Clinical psychologists are interested in identifying who will suffer psychological dysfunction but also who will not after an untoward early environment (e.g., exposure to violence);
- School administrators and staff want to identify students who are likely to engage in school shootings;
- Physicians and the rest of us want to identify who is at high risk for a particular type of cancer;
- The military is interested in who is likely to suffer PTSD or be a fabulous submarine commander;
- Coaches of professional football teams are keen to identify who will be the athlete (out of college) who is likely to be a great performer, especially under pressure; and

- Many are interested in identifying their soulmates and partners for life and separating them from creeps. (My "soul-mate matching service" is free to methodologists at www.kazdin-disharmony.com.)

All of those examples are classification examples. A goal of research is to identify the variables that help in selection and classification and to use that information to increase accuracy so that action can be taken as needed (e.g., for further diagnostic screening, for prevention). Key concepts are important to know for research purposes but also for one's personal life that relate to accuracy of classification. The concepts arise from epidemiology and public health where the observational designs especially flourish. Yet the designs and these particular concepts play an increasing role in psychological research.

> *The first term is sensitivity and refers to the rate or probability of identifying individuals who are predicted to show a particular characteristic (e.g., based on some screening measure or a set of variables) and in fact do show that outcome*. An easy and accurate way to refer to this as rate or percentage of identifying *true positives*.

That is, these are people who were identified (e.g., early in life) to show an outcome (e.g., disease) based on some assessment and in fact they actually do.

> *The second term is specificity and refers to rate or probability of identifying individuals who are not likely to show an outcome and in fact do not*. This refers to the rate or percentage of identifying *true negatives*.

For these individuals we said (based on our diagnosis and screening) that they would not show the problem later and we were right!

Sensitivity and specificity are probability statements that pertain to accuracies and inaccuracies in classification or identifying cases.

> The information for sensitivity and specificity often comes from the observational designs we have been discussing in this chapter. Clinical psychological research is interested in classification and case identification so the concepts are important.

I mentioned the concepts are important in everyday life as well. When a doctor says we or one of our loved ones is at risk for something horrible, she means that for a group of individuals with these characteristics (e.g., family history, something in our DNA or genetic code, type of skin) is likely to show or is at risk for some outcome. We would like to know how much risk because that can vary quantitatively from trivial to huge. Related, we would like to know more about sensitivity and specificity. That is, the predictions are probabilities and there will be misclassifications, including of course false positives (I said you would contract the problem but you did not) and false negatives (I said you would not get the problem, but you did—sorry).

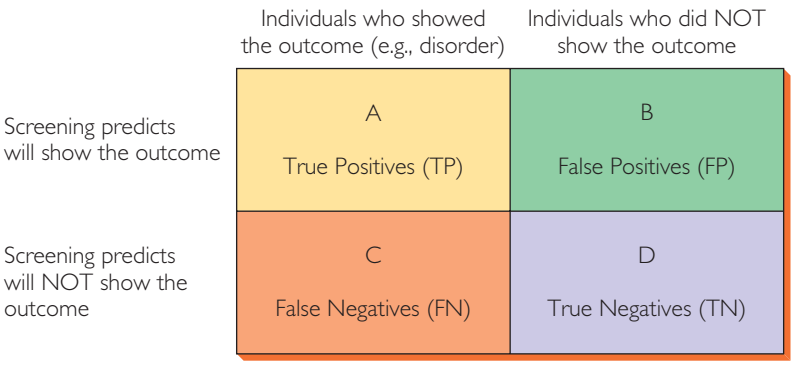
Some of the misclassification is due to errors of measurement. For example, for a psychology study, one might identify individuals who are depressed and use one or more measures and select a cutoff to operationalize depression. Some people with that score are not depressed and would not have that score on another day and some would have met that score on another day but did not. In psychological experiments, we are often interested in classification to carry out the observational designs I have reviewed already. Measures are rarely perfect or perfectly accurate and that can lead to misclassifications. (Later in this chapter, I mention the unresolved challenges in sports of classifying humans as male or female to decide who can participate in men's or women's track.) Yet misclassification also occurs simply because we do not know all the variables involved and their relative weight or contribution to making a prediction. Thus, we are simply in the dark. For example, not everyone who smokes heavily gets lung cancer but it is wise to tell a person he or she is at super risk but there will be some false positives—even though the measure (number of packs per days, number of years of smoking) is solid. And there are some false negatives—based on your history of never smoking, we said you were not likely to contract lung cancer but you did.

7.4.3: Further Considerations Regarding Sensitivity and Specificity

Sensitivity and specificity are about probabilities of accurately identifying individuals. Armed with these concepts, we can complete the full set of options and these are provided in Figure 7.3. It is useful to understand sensitivity and specificity. These concepts generate useful research studies as one can try to improve classification. A key issue is to understand that there can be huge trade-off in the cells noted in the Figure 7.3.

For example, if one wants to identify all the people who are terrorists at an airport, that is easy. Call everyone a terrorist. Do not even screen them—no need. If they show up at the airport, they go into Cell A (Figure 7.3). That will pick up those few terrorists and not miss any! Yet, the problem is clear, namely, that the false positive rate would be huge. Wait, we can get rid of false positives by classifying differently—we can say no one is a terrorist and, whew, we took care of that problem. Whoops—we missed identifying any of the terrorists. As we understand more about a phenomenon, we want to be able to increase accuracy of classification across the cells and to keep both sensitivity and specificity high. The trade-offs are not equal as we

Figure 7.3: Diagnosis or Classification of Individuals: Sensitivity and Specificity (Cell D)



Where

Sensitivity = TP/(TP + FN) or by using Cell identification = A/(A + C)
Specificity = TN/(TN + FP) or = D/(D + B)

NOTE: Although the formulas for computing the Cell values are noted here, the most important point to grasp is understanding of what sensitivity and specificity are and to be aware and wary of when one learns about some factor increasing risk for an outcome or accuracy of classification. In these cases, the data and formulas for computing are important.

understand more of the variables involved. That is, we can increase accuracy in classification while holding inaccuracy to a small percentage. To do this requires knowing more (about the variables involved) and being able to assess them reliably. These two tasks are major lines of research in mental and physical health (e.g., psychiatric diagnosis, responsiveness to treatment). (Figure 7.3 also provides additional terms used in evaluating sensitivity and specificity and computing the probabilities of interest. These are included for reference but are not elaborated further here.)

Observational research in clinical psychology relies on classification for selection of subjects for the designs we have discussed. Our research is mostly case-control designs in which classification is a beginning point to carry out the study. Typically, research does not focus on selection or diagnosis in the way that sensitivity and specificity analyses routinely do in public health and medicine. Yet psychological research is greatly interested in prediction and classification but usually goes about it slightly differently. For example, more common in psychology are regression analyses to identify variables and their weights (e.g., beta) in predicting an outcome or in delineating groups. Specificity and sensitivity analyses are another way to do this and provide valuable information about error rates in prediction. Yet, it is very important to be aware of sensitivity and specificity. The various permutations of classification (four Cells in Figure 7.2) are excellent sources of ideas for research; they are also important in everyday life as one makes decisions (e.g., about diet, surgery, mate selection through some matching service).

7.4.4: General Comments

I have mentioned the many benefits of prospective cohort designs ending that discussion with comments on classification. There are weaknesses of prospective longitudinal designs as well (see Table 7.3).

1. The design can take a considerable time to complete. Depending on the time frame (e.g., 5 or more years), the designs may not be well suited for addressing questions for which immediate or indeed urgent answers are needed (e.g., questions related to health, social policy, and welfare).
2. Longitudinal studies can be quite costly. The usual costs of personnel (research staff) are evident in any project, but longitudinal work may require additional costs of special personnel to remain in close contact with the subjects to retain them in the study and multiple payments to subjects and all participants (e.g., parents, teachers, children) who provide data or allow the project to go on.
3. If the study is conducted over an extended period (e.g., 2 or more years but perhaps up to 30 or 40 years), many cases can be lost over time (attrition). The potential for selection biases in the remaining sample and obstacles in estimating rates of the outcome are two of the problems that can emerge. The threat and likelihood of attrition are why very special attention is provided to the subjects, and project staff often are needed who are committed just to the retention of subjects. The special attention may include routine phone calls and letters, birthday and holiday cards, newsletters, and reminders about the project throughout the year just to keep the subjects interested or involved.

4. It is possible there will be cohort effects. That is something special about when the study began and was completed that may have made the results specific to the group (cohort) that was studied. This is not usually a major detriment in initiating a study but is something to be aware of when discussing the findings. This is a background external validity issue, namely, is there a good reason to believe the results will not generalize to another cohort?
5. The outcome of interest that one wishes to examine (e.g., onset of schizophrenia, criminal behavior) may have a low base rate in the population. Thus, if one is looking for the onset of criminal behavior, perhaps only 10% would be expected to show this in the cohort selected.

A sample of 100 cases (e.g., adolescents who seem at risk for criminal behavior) would not be sufficient for the data analyses because of the weak statistical power in detecting 10 cases in the at-risk group. If the 100 cases were divided into at-risk and not at-risk groups, there might be no difference in the outcome (criminal vs. no criminal) because of weak statistical power. A larger sample size is needed or cases need to be selected that are likely to have a higher base rate of the outcome of interest. This is why many studies in epidemiology and public health have large sample sizes and are population based (e.g., drawing large numbers from representative segments of the population). Representative samples are needed to get true incidence and prevalence in the populations, but the sheer number may be needed to detect phenomena whose outcomes are proportionately small (e.g., under 10% in the population).

7.5: Critical Issues in Designing and Interpreting Observational Studies

7.5 Identify the specific issues that the researcher needs to be aware of at the research design stage

I have not exhausted all of the variations of case-control and cohort designs (see Hulley, Cummings, Browner, Grady, & Newman, 2007). The variations that I have discussed are those most frequently used within psychology. More importantly, the designs convey the scope of questions that can be addressed. The challenge of the designs is isolating the construct of interest and the direction of influence among predictors and outcomes.

There are special issues that case-control and cohort studies raise to which the investigator ought to be particularly sensitive at the design stage. The issues pertain primarily to the construct validity of the findings, i.e., the extent to which the results can be attributed to the construct that the investigator wishes to study. Table 7.4 outlines several interrelated issues pertaining to construct validity.

Table 7.4: Critical Issues in Designing and Evaluating Case-Control and Cohort Studies

Critical Issues	Description
1. Specifying the Construct	• What is the construct of interest? • What are the operational criteria to separate or delineate groups (e.g., the specific measures or selection criteria)? • To what extent is the assessment procedure (e.g., criteria, measure) known to reliably separate or select persons with and without the characteristic?
2. Selecting Groups	• From what population, setting, or context (e.g., community, clinic) will the index sample be drawn? • If one group is to be compared with another that is selected at the outset of the study, what is this particular control or comparison group the one most suitable for the study? For what influences or constructs is it intended to control? • Are the groups with and without the characteristic of interest similar on subject and demographic variables (e.g., age, sex, race, socioeconomic status)? • Does the comparison group (without the characteristic) share all the characteristics but the one of interest? If not, how are these other characteristics to be evaluated, partialled out, or addressed in the design (e.g., additional control group[s] or data analyses)? • Could the construct as described (e.g., depression) be interpreted to reflect a broader construct (e.g., having a disturbance, being a patient)?
3. Direction and Type Influences	• Do the results permit conclusions about the time line, i.e., that one characteristic of the sample (e.g., exposure to an event, some experience) antedates the other? • Do the results permit conclusions about the role that one or more variables play in the outcome (i.e., risk factor, causal factor, mediator)?

7.6: Specifying the Construct

7.6 Express the importance of proper specification of the construct due to its impact on the findings

The first issue for the investigator is to specify the construct to study. As basic as this sounds, this can have tremendous implications for interpretation of the findings.

7.6.1: Level of Specificity of the Construct

Constructs that serve as the impetus for observational studies can vary in their level of specificity. Broad and global variables such as age, sex, social class, and ethnicity are less preferred as the basis of an investigation than more

specific variables with which these may be associated (e.g., patterns of interacting with friends, child-rearing practices, social support patterns). The more specific construct helps move from description of a relation (e.g., that males and females differ) toward explanation (e.g., those processes that may explain the differences).

To illustrate the point, consider for a moment that we are interested in studying the impact of SES on health. SES is a broad variable that encompasses (is related to) a plethora of other variables. SES has been studied extensively, and from this research we have learned that low SES (as measured by income, educational, and occupational status) predicts a very large number of untoward mental and physical health outcomes (e.g., higher rates of physical and mental illness, earlier death) (Adler, Bush, & Pantell, 2012; Aneshensel, Phelan, & Bierman, 2013; New York Academy of Sciences, 2010). This research has been extremely important.

A limitation of our knowledge is that we have not elaborated fully the reasons why these effects occur. The construct is very broad and encompasses so many other variables that we now need more specific studies to identify possible bases for the findings. Progress has been made in understanding some of the factors. For example, we know that schooling (amount of education) and income are two related mediating factors and that improving education and reducing poverty can improve health outcomes (e.g., Kawachi, Adler, & Dow, 2010). We also know that most of us as citizens have low mental health literacy, i.e., knowledge about what mental illness is, what can be done, and how to access services (Jorm, 2012). Yet, limited mental health literacy and actually access to care are associated with socioeconomic disadvantage. There is much more to the relation between low SES and poor health outcomes, but we have begun to identify some factors and places to intervene that actually can improve health, use of services, and clinical outcomes.

As a general guideline, broad constructs, such as SES, sex, and minority group status, often serve as a useful point of departure at the beginning of research. However, understanding is likely to be greatly enhanced by moving toward more specific constructs that might explain the processes through which the outcome might occur.

> On a continuum of description to explanation, research that can move toward the explanation side is usually more informative. In brief, specify the construct of interest and when possible hypothesize and test why the differences would occur.

7.6.2: Operationalizing the Construct

In a study where two or more groups are compared (e.g., depressed vs. not depressed), operationalizing the criteria to delineate groups raises important issues. What will be the specific criteria to delineate cases from controls? There are many separate issues. In the earlier discussion of single- and multiple-operationism, I noted that different measures may yield different groups. Thus, a self-report measure or clinical rating scale may be used to define individuals as cases in a case-control study. Among the questions, to what extent are the procedures, methods, and measures used to delineate groups valid and in keeping with prior findings? If possible within the design, it is desirable to have more than one operational definition that can be used to delineate groups.

In some areas of research, there have been single methods or measures that have been used to classify individuals. As examples, there are standard, single, and frequently used measures to assess depression (e.g., Beck Depression Inventory), marital satisfaction (e.g., Dyadic Adjustment Scale), adult psychopathology (e.g., Hopkins Symptom Checklist), child abuse (e.g., Child Abuse Potential Inventory), conflict and violence between marital partners (e.g., Conflict Tactics Scale), and many others. In these cases, a research tradition and literature have emerged in which one measure has become standard as a way of defining who is a case and who is a control (although these measures are often revised over time). On the one hand, the fact that one measure has been used in an area so extensively allows the findings to accumulate in a way that permits comparison and accretion of studies. On the other hand, one measure bears some risk, even when the measure is well investigated. The use of a single method of assessing the characteristic or problem (e.g., self-report, special format of the measure) may restrict generality of the conclusions across other modalities. For example, self-report of prejudice, alcohol consumption, or marital satisfaction may yield different results from other report or direct observation in the lab.

Regardless of what measure or operational criterion is invoked to classify subjects as cases or controls, we want to be sure that the measure is consistent and accurate in how individuals are classified. If the measure or criterion used to delineate the groups is unreliable, it could be that some of the individuals counted as "depressed" really ended up in the control group and some of the individuals not identified as depressed ended up in the case or index group. There would be a *diffusion* of the variable (internal validity threat) because both "cases" (individuals with the characteristic) and "controls" (individuals without the characteristic) were inadvertently in both groups instead of being restricted to their respective groups. The unreliability of measures often is surprising.

Among the dramatic examples, there has been keen interest in research in understanding racial, ethnic, and cultural differences because they can be critical moderators in both mental and physical health. Race has been used but with tremendous unreliability in classifying groups (e.g.,

European American, African American, Hispanic American) because there are no standard criteria and no firm biological classification system (Banton, 2010; Bernasconi, 2010; Gullickson & Morning, 2011). Among the many issues is that when investigators or subjects themselves identify race, the classification can be very unreliable. In considering the major or broad classifications of racial differences, obviously the unreliability within a study and across multiple studies will yield very inconsistent findings. The meaningfulness of the groups is easily challenged as well in part because of the enormous heterogeneity within a given group.

For many variables, reliability of classification does not seem to be a problem because groupings are obvious. ("Obvious" is a word that usually precipitates severe anxiety among methodologists—I have taken heavy medication just to write these next paragraphs.) For example, sex differences are a frequent source of research in biological and social sciences. Sex (being male or female) could be the most obvious classification variable before us (I just took some more medication), leaving aside the important issue of sexual identity. Sex is not so easily or perfectly classified because by many different measures, there are some males who will be classified as female and females who are classified as males. Visually looking at individuals to make the classification would not work perfectly (e.g., hermaphrodites). Hard core biological indices (e.g., chromosome composition and hormone levels) do not work perfectly either, at least with current measures (see Blackless et al., 2000).

A brief digression conveys the point in an interesting way. In the Olympic games and athletic competition more generally, there has been keen interest in classifying athletes as males or females, in large part to stop male competitors from entering women-only events. And in such sports as track and field, but many others, there are events for males and for females. Sorting people by sex to the correct locker rooms and events should be easy. ("Easy" in methodology means "really difficult" and "obvious" means "not clear at all.")

Sex testing was introduced into competitive sports in the 1960s after some interesting instances of athletes who competed in such events (e.g., one male who bound his genitals, entered as a woman named Dora and competed in the 1936 Olympics, the women's high jump—he placed fourth—just missing a medal). Methodologically speaking, assessment and classification are the issue—how to measure sex differences so that individuals can be placed into group to which they belong?

To address these issues, various methods of assessment were tried. Direct observation appears so scientifically sound that one forgets that the measure could be demeaning and discriminatory! For example, in the mid and late 1960s women were required to undress before a panel of doctors for international athletic competitions. This became "refined" by directly examining an athlete's genital region. At the 1968 Olympics, genetic testing was introduced as a less demeaning method (e.g., by analyzing for a sex chromatin as assessed by saliva; this was further modified in a later Olympics that allowed detection of a Y [so called male] chromosome gene). While such methods were less demeaning, they were hardly flawless.

For example, for chromosome testing, some women (~1 in 500 or 600) would show an abnormal result, not meet the definition of female, and could be disqualified. There are a number of disorders of sexual differentiation that could lead to aberrant (although quite reliable) laboratory results on a measure designed to differentiate sexes. Those disorders would make a female not meet some chromosome test but it would be an artifact of the disorder. For a few reasons, accuracy being one, the Olympic committee no longer screens for sex differences. Yet, the issue remains.

In the past decade, a young female world champion distance runner (Mokgadi Caster Semenya) from South Africa has won several world championship medals. Yet, these were not the usual victories. She sped past all other runners with wide margins, and her times often were so much faster than any obtained in previous women's events. This raised various suspicions including the possibility that she was using performance-enhancing drugs or other illicit substances or that she was not a female at all but really a male, or that she had a rare medical condition. She was tested, not allowed to participate in athletics for a while, and was part of a huge international issue as many others including political leaders and human rights individuals noted a racist theme that might underlie the scrutiny, insensitivity in how this was pursued, and violation of the runner's privacy (Cooky & Dworkin, 2013). Eventually, she was allowed to return to track and again won many races (e.g., a medal in 2012 Olympics). Is she "really" a female—yes. But if anyone asks that question in general, be sure to ask for the operational definition of male and female, i.e., precisely how will that be assessed? As the Olympic history on this matter shows, there are problems so far with obvious and not-so-obvious measures.

7.6.3: Further Considerations Regarding Operationalizing the Construct

There are critical political, social, and legal issues connected with classification of all kind in research (e.g., mental illness, ethnicity, who is "poor" and who is not) but also in everyday life (e.g., yes or no—is this the kind of person I want to be with for the rest of my life?). The methodological point pertains to the grouping itself, i.e., how the classification is made. In the sex difference, boys and girls and men and women can be distinguished for most purpose in

everyday life. Yet, for research that seeks reliable, replicable, and more objective means of making classifications, the classification is not perfect and merely looking at an individual (visual inspection) is not quite accurate apart from embarrassing and annoying. Also, genetic testing (given variation in genotype and phenotype) is considered not to be ready for prime time to help (Wonkam, Fieggen, & Ramesar, 2010). This discussion also merely refers to gross biological differentiation. When one adds to this gender identity or how one conceives of oneself, i.e., as more male or female, this becomes a more complex and has yet to be fully integrated into places where classification and individual rights to privacy are considered.

Let us move away from the example to a broader issue for your own research. When you select groups for a case-control study or read the report of others, raise the question—how are the groups delineated? On what measures? And why this way? (To state that prior research has done it this way is not usually a good answer unless your study is going to challenge the standard way. Saying you are following what others have done only means you have not thought about the issue and you are hoping that people before you did. In methodology, "hope" also is called "gambling.")

In making a classification, we usually rely on a particular measure (e.g., diagnostic instrument or more likely a particular scale or questionnaire). In cases where there may be unreliability of the measure, sometimes a large sample is assessed and only the extremes of the distribution are considered. For example, on some personality trait, one might assess a large group and for purposes of the study select those who are high (≥ 67th percentile) and compare them with those who are low (≤ 33rd percentile). The rationale is that it is the middle group that is likely to be more unreliably identified because a few points in one direction or the other could move them above or below the median. Selecting extreme groups can be very useful, depending on the goals of the study, but deleting a large segment of the sample (in our example, the middle third) can greatly distort the relations among the measures. The statistics that result (correlations, multiple correlations, beta weights, odds ratios) will be quite different from that would come from using the entire sample. The desirable practice here depends on the question. Sometimes one is only interested in talking about and studying a very special subgroup (e.g., extremely inhibited children) and focusing on a very special group is quite fine. Other times one wants to see the relation (e.g., correlation) across the entire spectrum (e.g., children who vary from inhibited to extraverted) and one includes all. The continuum can be divided (e.g., high, medium, low) on some characteristic for purposes of description but the full range to see the relation of one variable (e.g., depression) with another (e.g., later eating disorder).

7.7: Selecting Groups

7.7 Recognize the importance of selecting the right group in research

Identifying the construct and the means through which it will be assessed usually dictates the sample. Yet, it is useful to distinguish issues related specifically to the sample to draw attention to concerns that can undermine the inferences the investigator wishes to draw. The key question is to ask, what is the population from which the cases will be drawn? Among the options are samples from the community, clinic, or other social agency.

7.7.1: Special Features of the Sample

Cases that are drawn from a clinic or social agency may have special characteristics that make them unrepresentative of the larger community sample. As mentioned previously, these special characteristics may influence (moderate) the direction or magnitude of the relation between the variables of interest from what it would be like in the community sample. This is a particularly important point to underscore in psychology studies using case-control designs.

> In epidemiology, where case-control designs flourish, large-scale investigations often are completed that focus on representative and randomly selected cases.

For example, representative cases might be identified by sampling from multiple sites to represent the population of interest. Drawing from different geographical areas (e.g., of the country) and rural and urban settings or sampling across different countries would be examples. Once the areas are selected, random selection may be used by sampling randomly on the basis of streets, neighborhoods, or phone numbers. Invariably, such sampling (like the census) is not perfect (not everyone is home, has a telephone, or sends in printed measures), but the sample is clearly representative of the population within sampling error to the best that research can accomplish.

> In psychology's use of case-control and cohort designs, samples are often selected from special settings (e.g., clinics, agencies, schools) where some feature about the recruitment process may influence the associations that are studied.

For example, if one is interested, say, in studying agoraphobia (fear of open places) and in comparing cases versus controls, the population from which one samples may be critical. Individuals with agoraphobia who come to a clinic for treatment may be very special insofar as they have come to a clinic, by whatever means and that variable alone may contribute to or interact with the results. Perhaps they are more severely impaired (or less severely impaired because

they could leave their homes to begin with) or more (or less) likely to have additional (comorbid) disorders than individuals with agoraphobia in the community who never sought treatment. It is not necessarily the case that one sample is better than another—it depends on the question of the investigator. However, it is important to think about the population in a case-control or cohort study because features of that population may limit the construct validity conclusions of the study. That is, one cannot talk about the main characteristic of the sample (e.g., agoraphobia) without noting as well that it is patients who meet one or more other criteria such as self-selection or severity of dysfunction. Related here is external validity because one might not readily extend the conclusions to a nonself-selected group if a reasonable case can be made that they are likely to differ from those who were studied.

7.7.2: Selecting Suitable Controls

In case-control and two-group cohort studies, emphasis is given to defining the "case" group, i.e., those who have the characteristic or problem of interest. The control or comparison group warrants very careful consideration because it is often this group that limits the study and the kinds of statements the investigator can make. Typically, the investigator is interested in evaluating a special group (e.g., patients with bipolar disorder or schizophrenia, children with a specific disease, persons exposed to a special experience, people with interest in methodology) and wishes to make specific statements about this group on a set of dependent measures. The difficulty arises when that special group is compared with a "normal" (community sample) control group. This latter group includes persons who are identified because they do not have the disorder, dysfunction, or special experience.

> *Healthy controls often is the term used to refer to subjects who are from the community recruited because they do not meet the criteria for the dysfunction or disorder that is the main focus of the study.*

(As we will see, the very term "healthy controls" hints that there could be a construct validity problem.) The results invariably show that the special group (e.g., with bipolar disorder) is different from the healthy control subjects on the measures of interest (e.g., fMRI, some emotion regulation or cognitive task). Although the interpretation may focus on the special group (bipolar patients), the "healthy" comparison group is often insufficient to permit specific inferences to be drawn about the special group and the construct of interest.[1]

> Consider the following examples of studies where patients were compared with healthy control subjects:

- Bipolar adult patients show significant cognitive deficits (social cognitive domain and overmentalizing) compared with healthy controls (Montag et al., 2010);[2]
- Patients with schizophrenia differ in cortical thickness (portions of the brain) and in working memory with which cortical thickness is likely to be associated when compared with healthy controls (Ehrlich et al., 2012); and
- Patients with social phobia, when given a face-perception task (with emotional and neutral stimuli), show lower activation (fMRI) in areas of the brain related to emotional processing (precuneus and posterior cingulate regions) when compared with healthy controls (Gentili et al., 2009).

No cryptic or bulleted sentence can ever do justice to each of the studies that were cited. Yet the point can be made. In each of these studies, there is a construct validity problem. The authors want to say that the target group has unique characteristics and that those characteristics are related to the clinical disorder of interest. They may be completely correct, but we cannot tell from the studies. Differences between a patient group and a healthy control group could be due to being a patient, having impairment in any area of functioning associated with being a patient, and having a psychiatric disorder rather than the specific disorder included in the study. There are all sorts of other characteristics (e.g., physical symptoms and conditions, motor, perceptual, neuropsychological, and cognitive) that are associated with psychiatric disorders and any one of these could make a patient group differ from a nonpatient group. Also, many characteristics (e.g., genes, some symptoms) are general across many disorders (Caspi et al., 2014; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). In short, the construct validity problem emerges because the interpretations (construct) the authors would like to make are not really allowed by the design of the study. The findings raise the prospect that having a psychiatric disorder could explain the findings, leaving aside the specific disorder. Stated another way, the "healthy controls" allows the general conclusion such as "unhealthy" subjects differ from healthy controls. This level of generality is not what the investigators had in mind when designing the study.

7.7.3: Additional Information on Selecting Suitable Controls

The construct validity issue is clearer by looking at another set of studies that included patients and healthy controls, as in the studies mentioned previously. However, each of these studies added a third control group to reduce if not eliminate the construct validity concern I raised in the other studies. Consider another set of examples:

- Patients with bipolar disorder show greater impairment in neurological and cognitive functioning when compared to healthy controls. Yet, a third group was included, namely, individuals who did not have

bipolar disorder but did have mild cognitive impairment (Osher, Dobron, Belmaker, Bersudsky, & Dwolatzky, 2011). The finding showed that both patient groups were different from healthy controls but no different from each other on several measures (e.g., memory, executive functioning, information processing). They were different on other measures (visual-spatial processing, attention, and motor skills). The use of a patient control group clarifies the finding—some features were related to the specific disorder but many were not. Had only the healthy controls been included, all of the measures would have distinguished patients from nonpatients.

- Patients with a diagnosis of intermittent explosive disorder (IED) were higher on a measure of relational aggression (e.g., peer directed, romantic partner directed) than healthy controls (Murray-Close, Ostrov, Nelson, Crick, & Coccaro, 2010). (IED is characterized by extreme and uncontrollable expressions of anger that are disproportionate to events that seem to have precipitated them.) Yet, a third group was included of individuals who met diagnostic criteria for other disorders (varied). The patients with IED were much higher in relational aggression than both healthy control and other disorders groups. Inclusion of the other disorders group helped the construct validity because the authors can say having any diagnosis is not the reason relational aggression is high.
- Based on prior research, there was reason to expect brain differences (cortical thickness in various brain regions) for children with major depression (Fallucca et al., 2011). Magnetic resonance imaging was used, and children with depression were compared with healthy controls. Yet, a third group was included that consisted of children with obsessive-compulsive disorder—a patient group without the expectation of the brain characteristics evaluated in this study. The results found unique characteristics as predicted for the children with depression; obsessive-compulsive disorder and healthy controls were no different. The construct validity was greatly enhanced because we can rule out that the predicted difference would be evident by the presence of any disorder.

No one study is expected to be definitive and hence citing the individual studies as I did oversimplifies the area of work in which each study is conducted. Thus, each study I highlight might be one of several and across all of the studies, construct validity may be clearer. Even so, the construct validity issue I noted is important and undermines the findings of the first group of studies that did not include a comparison group other than healthy controls.

> If one wants to talk about a specific disorder in a case-control study, it is advisable to include another group with some other disorder not expected to show the characteristic. Healthy controls are fine to include but more often than not insufficient to support specific comments about the patient or target group.

In general, case-control studies require special efforts to isolate the construct of interest. Special attention is required in assessing the construct by making implausible other interpretations that might explain group differences. The selection of groups that vary in the construct of interest, and to the extent possible *only* in the construct of interest, is a critical beginning. Emphasis in developing a study is on identifying the case group, where much more attention must be given to deciding and finally selecting controls to which cases will be compared.

There is an interesting challenge in selecting healthy controls in clinical research on patient populations. Controls may be identified from a community sample and defined as individuals who do not meet criteria for the psychiatric diagnosis of interest or indeed for any diagnosis. Excluding individuals from the group of controls based on diagnostic information is fine; however, it is important to bear in mind that community samples have a significant amount of clinical dysfunction. Individuals sampled from the community, whether children, adolescents, or adults, show relatively high rates (~25%) of psychopathology (e.g., Kessler et al., 2004). Thus, sampling individuals from the community to serve as a control will inevitably include some individuals with clinical dysfunction. They may be weeded out once clinical dysfunction is assessed. I mention this in passing only to note that "healthy controls" are sort of super healthy. They do not show key characteristics of individuals in community samples, a quarter of whom are likely to show dysfunction. This may or may not be important to consider in the screening criteria used for controls. Again, it is important for the investigator to consider quite precisely what purpose the control group is to serve and to make sure that, to the extent possible, the selection criteria that are invoked address the specific issues the investigator has in mind.

7.7.4: Possible Confounds

A critical issue is that there may be variables that are possibly confounded with the selection criterion for delineating groups. For example, one might compare teenage mothers and female teenagers who do not have children. Any group differences on some set of dependent measures might be due to the differences in being a mother. Obviously, other variables may be different for these groups and are potential confounds that could explain

the results. For example, teen mothers tend to have lower SES, to drop out of school at higher rates, and to have previously experienced physical or sexual abuse, just to mention some features with which teen motherhood can be associated (e.g., Al-Sahab, Heifetz, Tamim, Bohr, & Connolly, 2012). Some effort has to be made within the study to address these other variables and their role in differentiating groups.

> If confounding variables are not evaluated, conclusions will be reached that the primary variable (motherhood status) was the basis of the findings. Yet, there are many plausible rival interpretations if key confounding variables are not considered in the design or data analysis.

Controlling for potential confounds is not a methodological nicety or added features to impress advisors or reviewers. The substantive conclusions of research depend on ruling out or making implausible threats to validity. Potential confounds are about threats to construct validity but also can lead one wildly astray if not controlled. How stark can differences be when controlling or not controlling confounds? Please do not sip your coffee for the next 2 minutes.

A recent study, the largest cohort study of its kind in the United States (N > 400,000) included adults (50–71 years of age) to evaluate whether coffee is related to an earlier-than-expected death (Freedman, Park, Abnet, Hollenbeck, & Sinha, 2012). The participants were followed for 14 years, and death was evaluated from diseases (cardiovascular, stroke but other causes such as accidents—combined indices of multiple outcomes, sort of an overall summary measure, are sometimes referred to as "all-cause mortality"). Main finding: The more coffee one consumed, the higher the rates of mortality; that is, there is a positive relation (more of one [coffee] is associated with more of the other [rates of mortality in the observed period]). This is the first conclusion, namely, higher rates of coffee drinking may not cause early death but it is definitely related. But as the paid TV commercials say, "Wait, there's more."

Coffee consumption is associated with (confounded by) higher rates of cigarette smoking and alcohol consumption, lower rates of exercise, and poor diet. Individuals who consume coffee are more likely to have these other characteristics. When these characteristics are controlled statistically, we have the second conclusion, i.e., coffee consumption and mortality are *inversely* related (more of one is associated with less of the other). Greater coffee consumption is associated with lower rates of early death. Controlling confounds led to the opposite conclusions. Very important to know because now one can look into how coffee may contribute—it does not seem to be caffeine only one of scores of compounds in coffee, because decaf had the same benefits!

As an aside, applying the findings to one's own life is interesting and both findings in the coffee example are relevant, namely, coffee is associated with a worse outcome (dying younger) or better outcome (dying older). Both are accurate. If one drinks a lot of coffee and also has some of the other characteristics (cigarette smoking, poor diet, etc.), the earlier death finding is more relevant. If one drinks a lot of coffee but does not have those other characteristics, the later death finding is more relevant.

> Another way to state all of this is to note that the impact of coffee consumption on early death is moderated by other health-related factors (cigarette smoking, poor diet, etc.). A moderator means that coffee consumption makes a contribution to outcome (not dying early) but the direction of its effect depends on a bunch of other things.

(Recall that one source of research ideas is the study of moderators. One can see from the coffee example how moderators can make a huge difference and in the process can be very interesting in the results they produce.)

7.7.5: More Information on Possible Confounds

Obviously, controlling confounds (or assessing moderators) can be critically important for the conclusions one reaches. There are several ways in which confounds can be addressed—some from the design of the experiment and some from the data analyses. From the standpoint of the design, groups (e.g., in a case-control study) can be matched on variables that could confound the group differences. For example, if the study compared teen mothers and female teenagers who do not have children one could match on SES, educational achievement, history of abuse, parent marital status (divorced, single), and family history of antisocial behavior, which are known to be related to early teen pregnancy. Mentioned before were more comprehensive ways of matching than just using a few or even hundreds of variables. Propensity score matching was mentioned as one set of ways in which this is done.

More commonly used are techniques in which potential confounding variables are entered into statistical tests (e.g., regression equations) as covariates. This latter method is a statistical way to ask—once these other variables are controlled, are female teens different in important ways as a function of having children?

What do you think?

One can match teen mothers and nonmothers on potentially confounding influences. A dilemma is that if groups are equalized or matched on such variables, the investigator cannot evaluate the impact of these variables in differentiating groups. Matching on a set of variables has to be decided

on the basis of the purpose of the study, i.e., whether one wishes to hold one variable constant so that others can be evaluated, or whether one wishes to identify the range of predictors that delineate groups. From the standpoint of the design, it is often useful to make the comparison with the confounds present (i.e., compare all the teen mothers and nonmothers) to see what the differences are. Then in the same study, it is useful to compare the mothers with a matched subsample (within the study) of nonmothers where the confounding differences (e.g., SES, education) are controlled. Thus, a comparison might be mothers with just those other nonmothers in the sample who are matched for SES and education or analyses with key variables (demographic variables that may not be of interest) controlled. This was how the study was done on coffee and mortality. Examine the relations (coffee consumption and death) with and without confounds controlled.

Data-analytic strategies play a major role in evaluating potential confounds. The goal of data analyses usually is to identify whether the variable of interest makes a contribution to the outcome independently of the confounding variable(s). The analyses can be done in many different ways. Statistical adjustments for possible confounding variables can be made (e.g., partial correlations, analyses of covariance) to consider confounding variables individually or as a group. Also, regression analyses can be completed (e.g., hierarchical regression, logistic regression) to test individual predictors (primary variable, confounding) in relation to the outcome.

> Statistical analyses (e.g., path analyses, structural equation modeling) can evaluate the relations in more integrative ways than just controlling or removing the impact. It is useful to precede statistical analyses with a conceptual model of the relation among variables that are being assessed. Conceptual models can specify the relations of constructs to each other (e.g., education, SES, abuse practices) and in relation to the outcome.

For example, in the hypothetical example of teen mothers versus females of the same age who are not mothers, the models can test whether education and SES make separate contributions to the outcome, whether their influence is direct or indirect (e.g., through some other variable), and the relative contribution (strength of the relations among different variables). Testing a model to evaluate multiple variables is an excellent way to handle potentially confounding variables. The reason is that "confound" is a relative concept, i.e., the main variable and potential confound in *my* study (e.g., SES and diet, respectively) may be the confound and main variable, respectively, in *your* study. If the issue is to understand multiple influences on an outcome and how they work together, use of a conceptual and statistical model to explain the interrelations among influences is an excellent design strategy.

7.8: Time Line and Causal Inferences

7.8 Determine how incorrect reporting of the predictor and the outcome leads to incorrect findings

A critical issue in case-control research pertains to the time line. One of the hazards the investigator must consider is to keep the conclusions in line with what the design can demonstrate. The most common problem is to imply a causal relation when the design does not permit comments about the time line. Consider as an example a cross-sectional, case-control study. The outcome of interest (grouping variable) may be an anxiety disorder in children (present or not) and the other characteristic (hypothesized antecedent) may be family stress. Children and their parents are assessed on a single occasion and complete various measures of child anxiety and family stress. The results may indicate that children who show the outcome (anxiety disorder cases), compared with those who do not (no-disorder controls), come from families that are more highly stressed. Clearly, the study demonstrates a correlation between two variables. The theory underlying the study may pose a directional relation in which family stress occurs before child dysfunction and through some process makes the child vulnerable, so that new stressors manifest themselves in anxiety. Actually, the results are consistent with hypotheses in either direction: stress as an antecedent to anxiety or anxiety as an antecedent to stress. In the absence of other evidence, this study does not establish stress as a risk factor for anxiety.

Statistical analyses commonly used in this type of research (e.g., discriminant analysis, logistic regression, structural equation modeling) may inadvertently contribute to the view that one variable precedes the other.

> The language of many data-analytic strategies identifies some variables as *predictors* or independent variables (e.g., family stress) and others as *outcomes* or dependent variables (e.g., presence or absence of anxiety disorder).

Also, computer output may have fancy lines and arrows to imply that one construct leads to another. The data analyses make no assumption of a time line for the variables that are entered; the distinction between antecedent (independent) and outcome (dependent), from the standpoint of the steps (discriminant function) of the analyses, is arbitrary. Clearly, the statistics are not at fault, but it is easy to misinterpret the results.

Consider how the language used in reporting results can exacerbate the misunderstanding. In our example, a typical conclusion might be worded that, family stress *predicted* child anxiety disorder (regression analysis, discriminant function) or family stress *increased the risk of* child anxiety disorder

(logistic regression). Such communications could be mistaken to suggest that family stress came first in the family stress-child anxiety sequence and even perhaps had a causal role in anxiety. "Predictor" in the output of a statistical program does not mean there is a timeline but in everyday life the word does. Understandably investigators complete their statistical analyses using the meaning of the statistical program and then in the discussion of the results slip into something more comfortable, namely, the implied time line.

7.9: General Comments

7.9 Report the utilities of case-controlled designs over experimentally studied ones

Case-control designs and their variations permit evaluation of human characteristics and experiences that usually cannot be readily studied experimentally. (One has to qualify this with "usually" because often nonhuman animal studies can vary the characteristics experimentally by randomly assigning to varied experiences or using techniques to induce a condition that serves as a model for what humans may have.) The designs are not inherently weak because they are observational, rather than experimental. Indeed, major advances in medicine, health, and nutrition, as just a few exemplary areas (e.g., risk factors for heart disease, various forms of cancer, impact of fats in one's diet) have emerged from such studies. The thinking and methodological sophistication of the investigator must be particularly astute with observational designs. Ingenuity of the investigator in selecting cases and controls and in data-analytic strategies that might be used to partial out influences is particularly important.

Most courses in methodology and research design in psychology do not include observational designs and their many options. This is unfortunate because the designs often are used in published research within clinical, counseling, and educational psychology. Of course, the main task of the investigator in observational or experimental research is essentially the same, namely, to decide in advance of the study precisely what he or she wishes to conclude. The precision of the statements one wishes to make determines key features of sampling, group formation, the design, and data analyses. In observational research, some threats to validity, such as subject selection (internal and external validity) and interpretation of the relation between the independent and dependent variables (construct validity) emerge in ways different from their equivalent in true experiments.

This chapter has focused on observational designs because of their frequent use in clinical research. The designs were treated at length to give attention to the many issues that can emerge in their execution and interpretation. It is important to note in passing that observational and experimental research can be combined in a single study. One might hypothesize that two groups of individuals (e.g., new criminal offenders vs. career criminals; or new methodologists vs. career methodologists) will respond differently to an experimental manipulation (e.g., a task that is designed to induce empathy). The study is both observational (cases, controls) and experimental (manipulation provided to one half of the cases and one half of the controls) and forms a 2×2 factorial design. Factorial designs are a convenient way to combine different types of variables and now in this context a way of combining different types of designs. I mention the designs again only to avoid the impression that research is *either* experimental or observational.

Summary and Conclusions: Case-Control and Cohort Designs

In observational studies, the investigator evaluates the variables of interest by selecting groups rather than experimentally manipulating the variable of interest. The goals of the research are to demonstrate associations among variables, but these associations may move beyond correlations to causal or at least approximations of causal relations. The studies can be descriptive and exploratory by trying to assess the scope of characteristics that may be associated with a particular problem or theoretically driven by trying to test models that explain the characteristics and how different influences relate to each other and to the outcome.

Case-control studies were identified and include those investigations in which groups that vary in the outcome or characteristic of interest are delineated. Typically two groups are compared (e.g., depressed vs. nondepressed patients) to evaluate a range of characteristics that may be evident currently (cross-sectional, case-control study) or may have occurred in the past (retrospective, case-control study).

> These designs are extremely valuable in understanding characteristics associated with a particular outcome, in unraveling the patterns of multiple influences and their relation, and in delineating subtypes by showing distinctions among individuals who have experienced the outcome (e.g., types of depression among the depressed group).

A limitation of these designs is that they do not permit strong influences to be drawn about what led to the outcome of interest.

Cohort studies are quite useful in delineating the time line, i.e., that some conditions are antecedent to and in fact predict occurrence of the outcome. In a single-group cohort design, a group that has not yet experienced the outcome of interest is assessed on multiple occasions and followed over time. At a later assessment, subgroups are delineated as those who do or do not show the outcome. Analyses can then identify what antecedents predicted the outcome. Birth-cohort studies have been a special case that have generated fascinating results related to physical and mental health because cases are often followed for decades.

Although a cohort study may begin with a single group, sometimes two or more groups are studied (multi-group, cohort design) to evaluate their outcomes. In this case, individuals may be selected because they show a characteristic but will be followed to examine yet another outcome. In some cases, multiple cohorts of different ages may begin the study and followed over time (accelerated, multi-cohort longitudinal design). The goal is to chart a particular developmental course over an extended period, but drawing on different groups to sample portions of that period.

Data from cohort studies often are used to classify, select, and predict a particular outcome. Sensitivity and specificity were discussed as key concepts related to the accurate identification of individuals who will show an outcome (sensitivity or true positives) as well as accurate identification of individuals who will not show an outcome (specificity or true negatives). The various permutations about predicting an outcome and in fact obtaining that outcome are all critical to understanding prediction in general and key areas of clinical research such as risk for a particular outcome.

> Case-control and cohort designs provide very powerful research strategies. The designs address a range of questions pertaining to how variables operate to produce an outcome (mediators, mechanisms) and the characteristics (moderators) that influence whether and for whom a particular outcome occurs.

The designs have been developed in other disciplines (epidemiology and public health) but are used routinely in clinical psychology. The designs require special attention to ensure construct validity of the results, i.e., that the conclusions can be attributed to the constructs the investigator has in mind, rather than to other influences. Critical issues related to designing and interpreting observational studies were discussed, including the importance of specifying the construct that will guide the study, selecting case and control groups, addressing possible confounds in the design and data analyses, and drawing causal inferences.

Critical Thinking Questions

1. What are the differences between true experiments and observational designs?
2. What are the differences between a concurrent, cross-sectional design and a prospective, longitudinal design?
3. What would be an example (hypothetical or real) of cohort design?

Chapter 7 Quiz: Observational Research: Case-Control and Cohort Designs

Chapter 8
Single-Case Experimental Research Designs

Learning Objectives

8.1 Identify some of the main features of the single-case experimental research designs

8.2 Analyze trend and variability as the two main aspects of stability of performance of an experiment

8.3 Report why experimental designs and their constituents are important in drawing the correct research conclusion

8.4 Describe the functionality of the ABAB design

8.5 Review the functionality of the multiple-baseline design

8.6 Determine the functionality of the changing-criterion design

8.7 Inspect how data is evaluated in single-case research

8.8 Analyze how visual inspection is a non-statistical method of evaluating single-case research

8.9 Express the causes that make statistical evaluation an important tool for evaluating single-case research

8.10 Scrutinize the strengths and the weaknesses of single-case designs

The *goal* of research is to draw scientifically valid inferences, i.e., research conclusions about phenomena that are as free as possible from the various threats to validity and sources of bias. The *means* consist of a variety of arrangements (designs) and practices (e.g., random assignment of participants, using various control groups, using reliable and valid measures, keeping observers naïve) to help achieve the goal. In considering methodology, researchers often focus primarily on methodological practices. For example, in group research certainly random assignment of subjects to conditions is an absolute must and to even question that violates key sections of the *Methodology Bible*. One reason we use random assignment is to distribute nuisance variables (all the potential influences we do not care about in our study) across groups, so the likelihood of selection biases is minimal. But sometimes we can achieve the goal in other ways (e.g., through various statistical matching procedures) and here too it is the goal rather than the practice that one might keep in mind. Mind you, I am not against random assignment—some of my best friends even use it. And after forceful demands of my dissertation proposal committee, I gave in and said I would use it myself. But sometimes other options are possible. These reminders of what we are doing and why, both of which are critical for methodological thinking in general and this chapter in particular.

Single-case experimental designs reflect an entire methodological approach to convey a broad set of such options that can achieve the goals of research but have a very different set of practices from those that are used in between-group research.[1] This chapter discusses ways of drawing causal influences and controlling threats to validity without necessarily using groups at all and indeed sometimes conducting experiments with one individual case. How can that be rigorous research? How can that be an experiment in which causal statements are made? Very easily as we shall see.

> The unique feature of single-case research designs is the capacity to conduct experimental investigations with the single case. As mentioned previously, the term "case" refers to an individual. Yet, this includes an individual person, a classroom, a school, a business, an entire city, or state. Case in this sense is the unit rather than necessarily number of people involved.

Single-case designs can evaluate the effects of interventions with large groups and address many of the questions posed in between-group research. However, the

methodology is distinguished by including an approach and multiple designs that rigorously evaluate interventions with one or a small number of cases. The designs have been used in experimental laboratory research with human and nonhuman animals and in applied research where interventions in the context of treatment, education, rehabilitation, and behavior change more generally (e.g., in business and community settings) are evaluated. The focus and illustration of this chapter are on the applied side that are direct relevant to clinical psychology, counseling, education, and other areas where there is an interest in changing something and identifying whether the intervention played a causal role. This chapter considers single-case experimental designs, including their underlying logic, how they address threats to validity, how they demonstrate causal relations, and many of the specific practices that allow them to yield scientifically valid inferences.

8.1: Key Requirements of the Designs

8.1 Identify some of the main features of the single-case experimental research designs

Single-case designs are true experiments. As we have used that term before, this means that the designs can demonstrate causal relations and can rule out or make implausible threats to validity. The underlying rationale of single-case experimental designs is similar to that of the more familiar group designs. All experiments compare the effects of different conditions (independent variables) on performance. In the more familiar between-group experimentation, the comparison is made between groups of subjects who are exposed to different conditions (experimental manipulations). In the simple case, after random assignment to conditions, some subjects are designated to receive a particular intervention and others are not. The effect of the manipulation is evaluated by comparing the performance of the different groups. In single-case research, inferences also are made about the effects of an intervention by comparing different conditions. Typically, these different conditions are presented to the same subject over time. Special characteristics of the design and various design options convey how this is done. We begin with the key requirements of single-case experiments that permit one to draw inferences about the effects of intervention. The requirements are summarized in Table 8.1 for easy reference.

8.1.1: Ongoing Assessment

The most fundamental design requirement of single-case experimentation is the reliance on repeated observations of performance over time.

> *Ongoing assessment consists of measuring the client's performance (e.g., direct observations, some other measure) repeatedly on several occasions, usually before the intervention is applied and continuously over the period while the intervention is in effect.*

Typically, observations are conducted on a daily basis or at least on multiple occasions each week. Ongoing assessment is a basic requirement because single-case designs examine the effects of interventions on performance over time. The assessment allows the investigator to examine the pattern and stability of performance before the intervention is initiated. The pre-intervention information over an extended period provides a picture of what performance is like without the intervention. When the intervention eventually is implemented, the observations are continued and the investigator can examine whether changes on the measure coincide with administration of the intervention.

The role of ongoing assessment in single-case research can be illustrated by examining a basic difference of between-group and single-case research. In both types of research, as already noted, the effects of a particular intervention on performance are examined, so comparisons can be made between when the intervention is presented and is not presented.

Table 8.1: Key Requirements of Single-Case Experimental Designs

Requirement	Definition	Purpose
Ongoing Assessment	Observations on multiple occasions over time prior to and during the period in which the intervention is administered	To provide the basic information on which data evaluation and intervention phases depend; decisions are made (e.g., when an intervention is effective or not, when to change phases in the designs) based on data derived from the assessment
Baseline Assessment	Assessment for a period of time prior to the intervention is implemented	To describe current performance and to predict what performance is likely to be like in the immediate future if the intervention were not implemented
Stability of Performance	Stable performance is one in which there is little or no systematic trend in performance and relatively little variability over time.	To permit projections of performance to the immediate future and to evaluate the impact of a subsequent intervention. A trend during baseline that is in the same direction as one hopes for which the intervention and highly variable performance (large fluctuations) can interfere with the evaluation.

In between-group research, the comparison is addressed by giving the intervention to some persons (intervention group) but not to others (no intervention or wait-list group). One or two observations (e.g., pre- and post-intervention assessment) are obtained for many different persons, and the comparison examines whether the groups differ.

In single-case research, the comparison usually is made with data from the individual or some small sets of individuals as the intervention is varied over time for those same individuals. There is no control group but there are controlled conditions, and it will become clear when we highlight the designs. Ongoing assessment refers to those several observations that are needed to make the comparison of interest with the individual subject.

A quick guide to help remember: Group research usually has *many subjects and few measurement occasions*; single-case research usually has *few subjects, but many measurement occasions*. There are exceptions to this but as a guide a good place to start.

8.1.2: Baseline Assessment

Usually each single-case experimental design begins with ongoing assessment for several days before the intervention is implemented.

> *This initial period of observation, referred to as the baseline phase, provides information about the level of behavior before a special intervention begins.*

The baseline phase serves two critical functions:

1. The first is referred to as *the descriptive function*. The data collected during the baseline phase describe the existing level of performance or the extent to which the client engages in the behavior or domain that is to be altered.
2. The second is referred to as *the predictive function*. The baseline data serve as the basis for predicting or projecting the level of performance for the immediate future if the intervention is not provided.

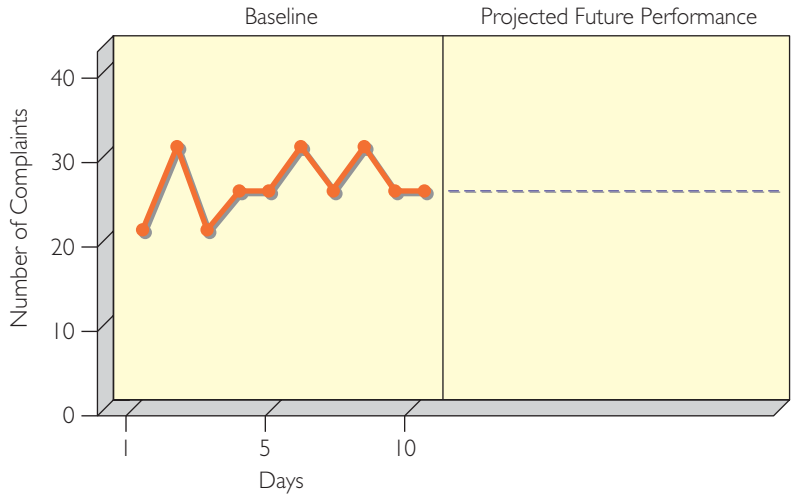
Of course, a description of present performance does not necessarily provide a statement of what performance would really be like in the future. Performance might change even without the intervention (e.g., from history or maturation, as just two possible influences). The only way to be certain of future performance without the intervention would be to continue baseline observations without implementing the intervention. This cannot be done because the purpose is to implement and evaluate the intervention in order to improve the client's functioning in some way. What can be done is to observe baseline performance for several days to provide a sufficient or reasonable basis for predicting future performance. The prediction is achieved by *projecting* or *extrapolating* a continuation of baseline performance into the future.

A hypothetical example can be used to illustrate how observations during the baseline phase are used to predict future performance and how this prediction is pivotal to drawing inferences about the effects of the intervention. Figure 8.1 illustrates a hypothetical case in which observations were collected on a child in a special education class and focused on frequency of shouting out complaints or comments to a teacher.

As evident in the figure, observations during the baseline (pre-intervention) phase were obtained for 10 days. The hypothetical baseline data suggest a reasonably consistent pattern of shouting out complaints each day in the classroom.

Figure 8.1: Hypothetical Example of Baseline Observations of Frequency of Complaining

The data in baseline (solid line) are used to predict the likely rate of performance in the future (dashed line).



We do not really know what performance will be like on days 11, 12, and so on—all those days after baseline that were not yet observed. Yet, the baseline level can be used to project the likely level of performance in the immediate future if conditions continue as they are. The projected (dashed) line predicts the approximate level of future performance.

This projected level is essential for single-case experimentation because it serves as one criterion to evaluate whether the intervention leads to change. Presumably, if the intervention is effective, performance will be different from the predicted level of baseline.

For example, if a program is designed to reduce shouting and is successful in doing so, the projected line (data points) for shouting out should be well below the projected line that represents the level of baseline. In any case, ongoing assessment in the beginning of single-case experimental designs consists of observation of baseline or pre-intervention performance.

8.2: Stability of Performance

8.2 Analyze trend and variability as the two main aspects of stability of performance of an experiment

Since baseline performance is used to predict how the client will behave in the future, it is important that the data are stable. A *stable rate* of performance is characterized by the absence of a trend in the data and relatively little variability in performance. The notions of trend and variability are the two facets of stability.

8.2.1: Trend in the Data

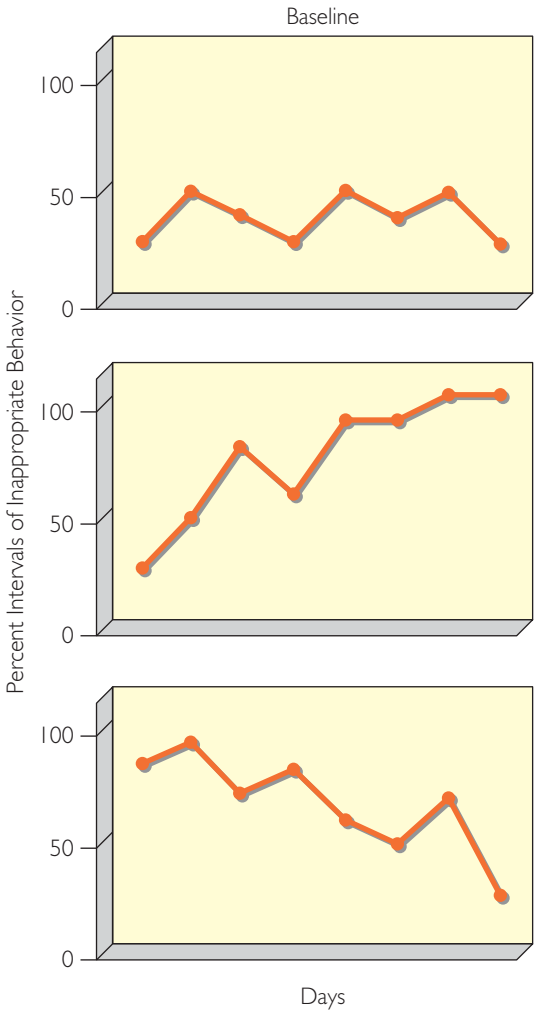
A trend line or slope refers to the line on a graph showing the general direction of the data points, i.e., in what direction they are heading.

I will use the term trend as an abbreviation for trend line. Consider just linear trend lines in which the data tend to be going up (accelerating) or down (decelerating) over time or is just flat and not going in either direction over time. This gradient or how steep the straight line is referred to as slope.[2]

One of three simple data patterns might be evident during baseline observations. First, baseline data may show no accelerating or deceleration pattern. In this case, performance is best represented by a horizontal or flat line indicating that it is not increasing or decreasing over time. As a hypothetical example, consider observations of a child's inappropriate and disruptive behavior (e.g., rarely in his seat, disrupts, handles the work of others while they are working, and blurts out comments during class). The upper panel of Figure 8.2 shows baseline performance with a trend line that is not accelerating or decelerating. The trend line is straight and generally horizontal over time. The absence of an accelerating trend in the baseline provides a relatively clear basis for evaluating subsequent intervention effects. Improvements in performance are likely to be reflected in a decelerating trend line (e.g., decreasing disruptive behavior) that departs from the horizontal line of baseline performance.

Figure 8.2: Hypothetical Data for Disruptive Behavior of a Hyperactive Child

The *upper panel* shows a stable rate of performance with no systematic trend over time. The *middle panel* shows a systematic trend with behavior becoming worse over time. The *lower panel* shows a systematic trend with behavior becoming better over time.



This latter pattern of data (lower panel) is the most likely one to interfere with evaluation of interventions, because the change is in the same direction as of change anticipated with the intervention.

An accelerating or decelerating trend line during baseline may or may not present problems for evaluating intervention effects, depending on the direction of the

trend in relation to the desired change in behavior. Performance may be changing in the direction opposite from the direction of change the intervention is designed to achieve.

For example, our child with disruptive behavior may show an increase in these behaviors (getting worse) during baseline observations. The middle panel of Figure 8.2 shows how baseline data might appear; over the period of observations, the child's behavior is becoming worse, i.e., more disruptive. Because the intervention will attempt to alter behavior in the opposite direction, i.e., improve behavior, this initial trend will interfere with evaluating intervention effects.

In contrast, the baseline trend may be in the same direction that the intervention is likely to produce. Essentially, the baseline phase may show improvements in behavior. For example, the behavior of the child may improve over the course of baseline as disruptive an inappropriate behaviors decrease, as shown in the lower panel of Figure 8.2. Because the intervention attempts to improve performance, it may be difficult to evaluate the effect of the subsequent intervention. The projected level of performance for baseline is toward improvement. A very strong intervention effect would be needed to show clearly that the change surpassed and departed from this projected level of baseline performance.

If baseline is showing an improvement, one might raise the question of why an intervention should be provided at all. Yet even when behavior or some measure is improving during baseline, it may not be improving quickly enough or not be close to some final goal. For example, infant mortality, crime rates, child and partner abuse, and car accidents might be declining (improving) in a given city, state, or country. Yet it is likely there will be interest in accelerating the decline if at all possible. Hence, even though behavior is changing in the desired direction, additional changes in the same direction may be needed or needed more quickly than just waiting it out. Perhaps the magnitude of the change (a sharp change in the trend line) will make the demonstration clear.

For present purposes, it is important to convey that the one feature of a stable baseline is little or no trend line in the same direction that is expected to occur during the intervention phase.

A pattern with a trend that is horizontal or in the opposite direction (e.g., accelerating) from what is expected or hoped for during the intervention (e.g., decelerating) provides a clear basis for evaluating intervention effects.

Presumably, when the intervention is implemented, a trend toward improvement in behavior will be evident. This is readily detected with an initial baseline that does not already show some improvement just as a function of time.

8.2.2: Variability in the Data

In addition to trend line, stability refers to the fluctuation or variability in the subject's performance over time. Excessive variability in the data during baseline or other phases can interfere with drawing conclusions about the intervention.

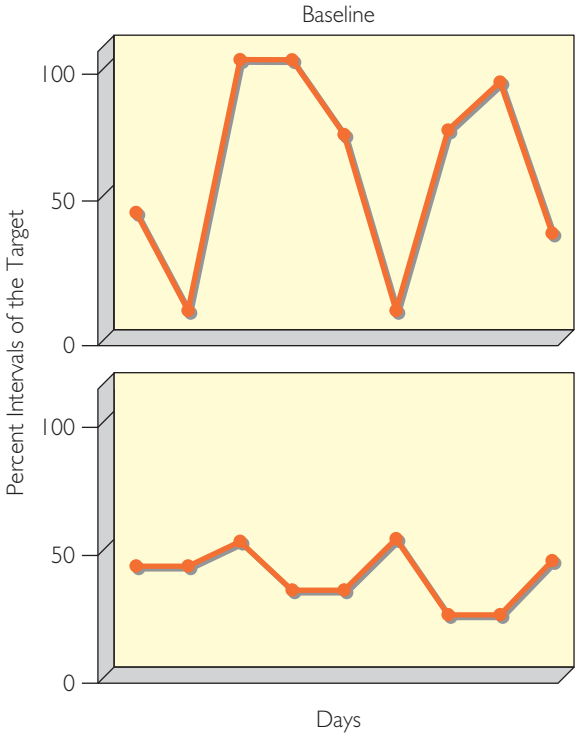
As a general rule, the greater the variability in the data, the more difficult it is to draw conclusions about the effects of the intervention.

And that general rule applies to group research, as reflected in our discussion of how excessive or extraneous variability can interfere with showing the effect of an intervention and is a threat to data-evaluation validity.

Excessive variability is a relative notion. Whether the variability is excessive and interferes with drawing conclusions about the intervention depends on many factors, such as the initial level or rate of behavior during the baseline phase and the magnitude of behavior change when the intervention is implemented. In the extreme case, baseline performance may fluctuate daily from extremely high to extremely low levels (e.g., 0% to 100%). Such a pattern of performance is illustrated in Figure 8.3 (upper panel), in which hypothetical baseline data are provided.

Figure 8.3: Baseline Data Showing Relatively Large Variability and Relatively Small Variability

Baseline data showing relatively large variability (*upper panel*) and relatively small variability (*lower panel*). Intervention effects are more readily evaluated with little variability in the data.



With such extreme fluctuations in performance, it is difficult to predict any particular level of future performance. Alternatively, baseline data may show relatively little variability. As represented in the hypothetical data in the lower panel of Figure 8.3, performance fluctuates, but the extent of the fluctuation is small compared with the upper panel. With relatively slight fluctuations, the projected pattern of future performance is relatively clear and hence intervention effects will be less difficult to evaluate.

Sometimes there is no variability in performance during baseline because the behavior never occurs (e.g., exercising at home or at a gym, taking one's medication, practicing a musical instrument) or occurs every time or almost every time (e.g., constant complaining, tantrums, or swearing, having dessert with dinner). Consider the first scenario in which the behavior one wishes to develop does not occur at all before the intervention. The baseline observations might show zero occurrences each day and of course no variability. This will make intervention effects relatively easy to evaluate because any improvement (sometimes the behavior occurs) will be easily identified given the baseline rate (the behavior never occurs).

Variability can result from all sorts of influences—the behavior may be variable on its own. That is, as humans our performance fluctuates routinely and not being perfectly consistent or the same every day is the rule rather than the exception. Also, conditions of assessment can be loose or highly variable so that observers or the circumstances (e.g., activities, time of the day) surrounding the assessment contribute to fluctuations. Conditions of assessment often are held constant, and reliable measures (e.g., direct observations with trained observers, automated measures, measures that are relatively immune to rater biases) are used to be sure that little variability is due to the procedures of the study. This point is true of between-group designs too, i.e., we choose reliable and valid measures in part to be sure excessive variability (error) does not detract from detecting an effect when there is one.

8.3: Major Experimental Design Strategies

8.3 Report why experimental designs and their constituents are important in drawing the correct research conclusion

The key requirements are basic ingredients of single-case designs and provide the information that is used to draw inferences about intervention effects. Yet, by themselves they do not permit one to draw causal inferences about the impact of the intervention. For that, we need the experimental designs, i.e., how conditions are presented and evaluated over time. Major designs are presented and illustrated here. (See Further Readings for sources with additional design options.)

8.4: ABAB Designs

8.4 Describe the functionality of the ABAB design

ABAB designs consist of a family of experimental arrangements in which observations of performance are made over time for a given client (or group of clients).

8.4.1: Description

The basic design variation examines the effects of an intervention by alternating:

- The baseline condition (A phase), when no intervention is in effect
- The intervention condition (B phase)

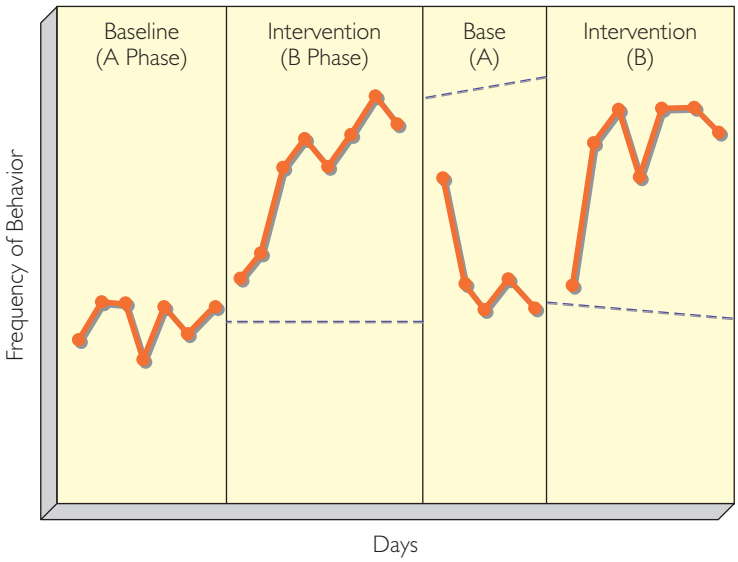
The A and B phases are repeated again to complete the four phases. The effects of the intervention are clear if performance improves during the first intervention phase, reverts to or approaches original baseline levels of performance when the intervention is withdrawn, and improves when the intervention is reinstated in the second intervention phase.

FIRST PHASE (BASELINE) The design begins with baseline observations when behavior is observed under conditions before the intervention is implemented. This phase is continued until the rate of the response appears to be stable or until it is evident that the response does not improve over time. As mentioned previously, baseline observations serve two purposes, namely, to describe the current level of behavior and to predict what behavior would be like in the future if no intervention were implemented. The description of behavior before the intervention obviously is necessary to give the investigator an idea of the nature of the problem. From the standpoint of the design, the crucial feature of baseline is the prediction of behavior in the future. A stable rate of behavior is needed to project what behavior would probably be like in the immediate future. Figure 8.4 shows hypothetical data for an ABAB design. During baseline, the level of behavior is assessed (solid line), and this line is projected to predict the level of behavior into the future (dashed line). When a projection can be made with some degree of confidence, the intervention (B) phase is implemented.

SECOND PHASE (INTERVENTION) The intervention phase has similar purposes to the baseline phase, namely, to describe current performance and to predict performance in the future if conditions were unchanged. However, there is a third or added purpose of the intervention

Figure 8.4: Hypothetical Data for an ABAB Design

The solid lines in each phase reflect the actual data. The dashed lines indicate the projection or predicted level of performance from the previous phase.



phase, namely, to test a prior prediction. Here is how the test part works. In the baseline phase, a prediction was made about future performance if baseline, instead of the intervention, were continued. In the intervention phase, the investigator can test whether performance during the intervention phase (phase B, solid line) actually departs from the projected level of baseline (phase B, dashed line). In effect, baseline observations were used to make a prediction about performance and data during the first intervention phase can test the prediction.

Do the data during the intervention phase depart from the projected level of baseline?

What do you think?

If the answer is yes, this shows that there is a change in performance. In Figure 8.4, it is clear that performance changed during the first intervention phase. At this point in the design, it is not entirely clear that the intervention was responsible for change. Some other influences coincident with the onset of the intervention might have caused the change (e.g., history as a threat to internal validity or maturation if the change was slow rather than abrupt). We know we have a change, but establishing the likely cause of that change requires more.

THIRD PHASE In the third phase (the second A of ABAB), the intervention is usually withdrawn and the conditions of baseline are restored. The second A phase describes current performance and predicts what performance would be like in the future if this second A phase were continued. There is a third purpose of the second A phase and of any phase that repeats a prior phase. The first A phase made a prediction of what performance would be like in the future (the dashed line in the first B phase). This was the first prediction in the design, and like any prediction, it may be incorrect. The second A phase restores the conditions of baseline and can test the first prediction. If behavior had continued without an intervention, would it have continued at the same level as the original baseline or would it have changed markedly? The second A phase examines whether performance would have been at or near the level predicted originally. A comparison of the solid line of the second A phase with the dashed line of the first B phase in Figure 8.4 shows that the lines really are no different. Thus, performance predicted by the original baseline phase was generally accurate. Performance would have remained at this level without the intervention.

FINAL PHASE In the final phase of the ABAB design, the intervention is reinstated again. This phase serves the same purposes as the previous phase, namely, to describe performance, to test whether performance departs from the projected level of the previous phase, and to test whether performance is the same as predicted from the previous intervention phase. (If additional phases were added to the design, the purpose of the second B phase would of course be to predict future performance.)

> In short, the logic of the ABAB design and its variations consists of making and testing predictions about performance under different conditions. Essentially, data in the separate phases provide information about present performance, predict the probable level of future performance, and test the extent to which predictions of performance from previous phases were accurate.

By repeatedly altering experimental conditions in the design, there are several opportunities to compare phases

and to test whether performance is altered by the intervention. If behavior changes when the intervention is introduced, reverts to or near baseline levels after the intervention is withdrawn, and again improves when the intervention is reinstated, then the pattern of results suggests rather strongly that the intervention was responsible for change. All sorts of other influences that might be history or maturation (e.g., coincidental changes in the behavior of parents, teachers, spouses, an annoying peer, bosses at work, or in internal states of the individual such as change in medication, onset of a worsening of cold, broken smart phone, unfriended by all one's relatives on some social media platform, new social relationships) that might be responsible for behavior change are not very plausible in explaining the pattern of data across phases and the replication of intervention effects. Other threats to internal validity such as testing or statistical regression too cannot explain the data pattern. The most plausible explanation is that the intervention and its withdrawal accounted for changes.

8.4.2: Illustration

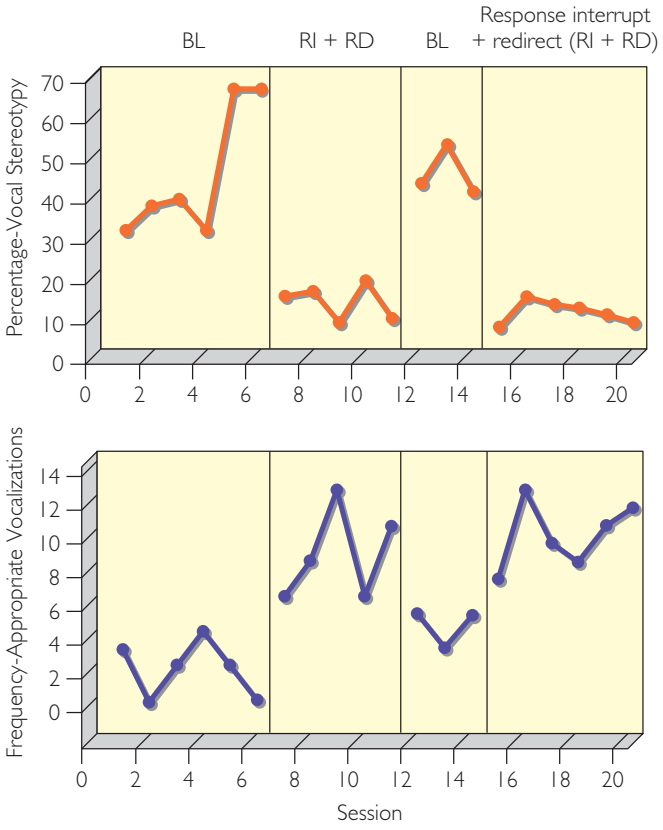
The ABAB design was used to evaluate an intervention to reduce vocal stereotype among children diagnosed with autism spectrum disorder and who were referred because their vocalizations interfered with their participation in other special educational activities (Ahearn, Clark, MacDonald, & Chung, 2007). Vocal stereotype refers to vocalizations such as singing, babbling, repetitive grunts, squeals, and other phrases (e.g., "ee, ee, ee, ee") that are not related to contextual cues in the situation and appear to serve no communication function as part of interaction with others. Individual sessions were conducted, 5 minutes in duration, in which a child sat in the room with the teacher. Both stereotypic and appropriate vocalizations (e.g., "I want a tickle," "Could I have a chip?") that communicated content were recorded.

Baseline (no intervention) was followed with an intervention phase that included response interruption and redirection. This consisted of immediately interrupting any vocal stereotype statement and redirecting the child to other vocalizations. The teacher would state the child's name and then ask a question that required an appropriate response (e.g., "What is your name?" "What color is your shirt?"). Any spontaneous appropriate verbal statement was praised (e.g., "Super job talking!"). Observations were obtained across several intervals to score the presence or absence of the stereotypic and appropriate vocalizations. The response interruption and redirection intervention was evaluated in an ABAB design. Figure 8.5 provides data for one of the children, a 3-year-old boy named Mitch.

The data are clear. Stereotypic sounds (top graph) decreased and appropriate vocalizations (bottom graph) changed markedly whenever the response interruption and redirection intervention was implemented.

Figure 8.5: Evaluation of Response Interruption and Redirection Intervention

The percentage of each session with stereotypic behavior (top) and appropriate speech (bottom).



This was a demonstration of effects in a 5-minute controlled period and by itself is hardly something that might exert impact on the lives of the children. Often initial demonstrations are done exactly like this to evaluate whether a specific intervention can have impact. (This is analogous to a proof of concept discussed previously, i.e., to show what can happen in artificial and controlled circumstances.)

> Such controlled circumstances can be useful to identify what among various alternative interventions will work. Once an effective intervention is identified, it can be extended to everyday settings.

Indeed, in this study, after the demonstration, the intervention was extended to the everyday classroom and the benefits were evident there as well with no further use of a reversal phase.

The illustration conveys how the ABAB designs achieve the goals of ruling out or making threats to validity implausible. The changes when the phase was shifted from no intervention (A) to intervention (B) and back and forth again make the intervention the most plausible explanation for what led to the change. If one invoked the logic of single-case designs (describe, predict, and test a prior prediction), then the interventions are very likely to be the

reason for the change. There is no certainty in science from any single empirical test, but the preceding illustration is a strong demonstration of intervention effects.

8.4.3: Design Variations

There are many variations of the design based on several features. The number of phases can vary. The minimal configuration for an experimental evaluation is an ABA (three phases) design that may be:

- Baseline
- Intervention
- Baseline phases
- BAB (intervention, baseline, intervention)

That is the minimum because two phases that predict the same performance (e.g., baseline and return to baseline) are needed to show a replication of the effect. Obviously, four (or more) phases provide a stronger basis for drawing inferences about the impact of the intervention. An AB (two-phase) version is not usually considered to be an experimental demonstration because the description, prediction, and test of predictions logic cannot be invoked.

Variations of the design also include applications of different interventions. The ABAB version includes a one intervention (in the B phase) that is given to the client in two phases. Yet, sometimes the intervention is not effective or not sufficiently effective to achieve the intervention goals. A second intervention or variation of the first intervention (B_2 phase) might be added. This might be summarized as $AB_1B_2A\ B_2$. Here B_1 was not very effective, so the investigator changed the intervention B_2. That intervention was effective and now is included in the rest of the design. Is B_2 responsible for change? (Or as Hamlet asked in Shakespeare's play of the same name, "2B or not 2B?") The effect is replicated by presenting and withdrawing the intervention as per the requirements of an ABAB design.

> An important feature of ongoing assessment in single-case research is the ability to see whether behavior is changing and changing sufficiently. If it is not, the intervention can be modified (B_2, B_3) as needed to improve the effects achieved for the client.

ABAB designs also can vary by precisely what is implemented in the reversal (2nd A) phase. The most commonly used alternative in the reversal phase is to *withdraw the intervention*. This usually restores the conditions that were in place during the baseline (pre-intervention) phase. Understandably, behavior would likely revert to baseline levels when the intervention is withdrawn. Occasionally, a critical ingredient of the intervention, rather than the entire intervention, is omitted that also might lead to a return to baseline levels. There are many other variations of the design (see Kazdin, 2011).

8.4.4: Considerations in Using the Designs

ABAB design nicely illustrates the underlying basis of experimental research by showing how one can draw conclusions by isolating the effect of the intervention. When the changes in behavior follow changes in phases, as illustrated previously, this is a very strong demonstration that the intervention was responsible for change. Several issues emerge in using this design.

The design requires that behavior reverts to or approaches the original baseline level after the intervention is withdrawn or altered (during the second A phase). This requirement restricts the use of the design in many applied settings such as schools or the home in contrast to, for example, basic laboratory research (e.g., nonhuman animal research) where there is no applied goal. Educators (and others) obviously want the benefits the intervention to continue, i.e., not to revert to baseline levels. Thus, from an applied standpoint, continued performance of the appropriate behavior is important and desirable. Yet, from the standpoint of an ABAB design, it could be disappointing if behavior is not made to revert to baseline levels after showing an initial change. Without such a reversal, it is not clear that the intervention was responsible for the change.

Essentially, returning the student or client to baseline levels of performance amounts to making behavior worse. Of course, the intervention can be withdrawn for only a brief period such as 1 or a few days (e.g., Brooks, Todd, Tofflemoyer, & Horner, 2003; Wehby & Hollahan, 2000). Yet, in most circumstances, the idea of making a client worse just when the intervention may be having an effect is ethically unacceptable. Aside from ethical problems, there are practical problems as well. It is often difficult to ensure that the teacher, parent, or other agent responsible for conducting the program will actually stop the intervention during the return-to-baseline phase once some success has been achieved. Even if they do stop the intervention, behavior does not always revert to baseline levels.

As a general rule:

1. If a reversal *does occur* as conditions are returned to baseline that may be problematic if the behavior is important for the clients or for those in contact with them.
2. If a reversal *does not occur*, this raises obstacles in concluding that the intervention led to the change. Yet the power of the design in demonstrating control of an intervention over behavior is very compelling.

If behavior can, in effect, be "turned on and off" as a function of the intervention, this is a potent demonstration of a causal relation. There are several solutions that allow use of ABAB designs even when this might seem undesirable to reverse behavior. Among the options is to use special

procedures in the final B phase that are specifically designed to maintain behavior (see Kazdin, 2013a). But from a methodological standpoint, other designs are readily available that demonstrate a causal relation without using a reversal of conditions.

8.5: Multiple-Baseline Designs

8.5 Review the functionality of the multiple-baseline design

With multiple-baseline designs, the effects are demonstrated by introducing the intervention to different baselines (e.g., behaviors or persons) at different points in time.

8.5.1: Description

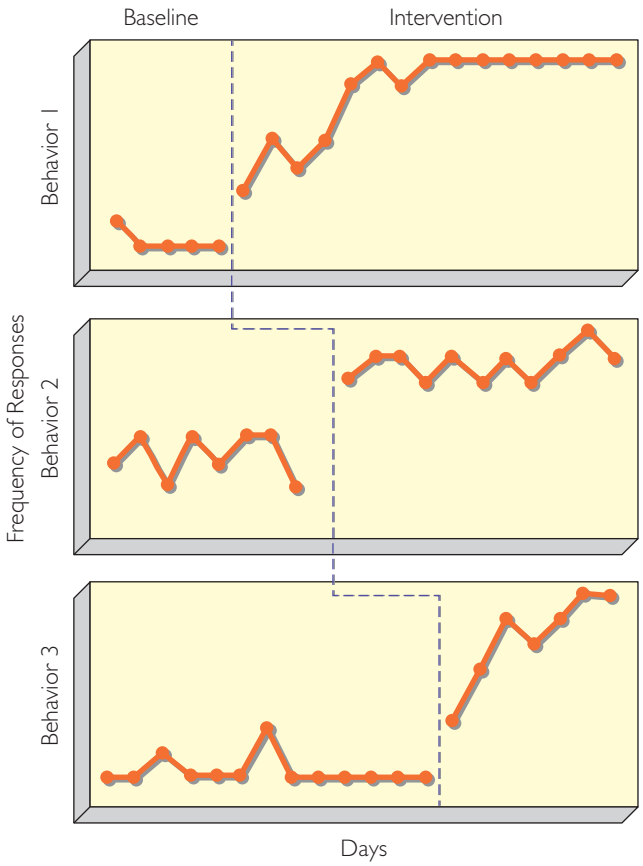
In multiple-baseline designs, if each baseline changes when the intervention is introduced, the effects can be attributed to the intervention rather than to extraneous events. Once the intervention is implemented to alter a particular behavior, it need not be withdrawn. Thus, within the design, there is no need to return behavior to or near baseline levels of performance as was the case with ABAB designs.

Consider the *multiple-baseline design across behaviors*, a commonly used variation in which the different baselines refer to many different behaviors of a particular person or group of persons. Baseline data are gathered on two or more behaviors. Figure 8.6 plots data from a hypothetical example in which three separate behaviors are observed. The baseline data gathered on each of the behaviors serve the purposes common to each single-case design, namely, to *describe* the current level of performance and to *predict* future performance. After performance is stable for all of the behaviors, the intervention is applied to the first behavior. Data continue to be gathered for each behavior. If the intervention is effective, one would expect changes in the behavior to which the intervention is applied. On the other hand, the behaviors that have yet to receive the intervention should remain at baseline levels. After all, no intervention was implemented to alter these behaviors. When the first behavior changes and the others remain at their baseline levels, this suggests that the intervention may have been responsible for the change but more is needed to make this more plausible.

After performance stabilizes across all behaviors, the intervention is applied to the second behavior. At this point, both the first and second behavior are receiving the intervention, and data continue to be gathered for all behaviors. As evident in Figure 8.6, the second behavior in this hypothetical example also improved when the intervention was introduced.

Figure 8.6: Hypothetical Data for a Multiple-Baseline Design across Behaviors

Hypothetical data for a multiple-baseline design across behaviors in which the intervention was introduced to three behaviors at different points in time.



Finally, after continuing observation of all behaviors, the intervention is applied to the final behavior, which changed when the intervention was introduced.

The design demonstrates the effect of an intervention by showing that behavior changes when and only when the intervention is applied. The pattern of data in Figure 8.6 argues strongly that the intervention, rather than some extraneous event, was responsible for change. Extraneous factors might have influenced performance. For example, it is possible that some event at home, school, or work coincided with the onset of the intervention. Yet one would not expect this extraneous influence to alter only one of the behaviors and at the exact point that the intervention was applied. A coincidence of this sort is possible, so the intervention is applied at different points in time to two or more behaviors. The pattern of results illustrates that whenever the intervention is applied, behavior changes. The repeated demonstration that behavior changes in response to staggered applications of the intervention usually makes the influence of extraneous factors implausible.

As in the ABAB designs, the multiple-baseline designs are based on testing of predictions. Each time the intervention

is introduced, a test is made between the level of performance during the intervention and the projected level of the previous baseline. Essentially, each behavior is a "mini" AB experiment that tests a prediction of the projected baseline performance and whether performance continues at the same level after the intervention is applied.

There are two added features that make this not just an AB design. Intervention is staggered in its presentation to different behaviors, and one can look for a pattern across all of the behaviors. If the intervention is responsible for change, one predicts no change on other baselines until the intervention is applied. Predicting and testing of predictions over time is similar in principle for ABAB and multiple-baseline designs, although carried out slightly differently.

8.5.2: Illustration

A multiple-baseline across individuals is illustrated in a program designed to alter the behavior of three African-American students (ages 8–10) in a special education classroom composed of eight students (Musser, Bray, Kehle, & Jenson, 2001).

The three students met criteria for two psychiatric disorders, namely:

- Oppositional defiant disorder (extremes of stubbornness, noncompliance)
- Attention deficit hyperactivity disorder (inattention, hyperactivity)

These are sometimes referred to as disruptive behavior disorders because the behaviors "disrupt" others and the environment. The focus was on reducing disruptive behaviors in class (e.g., talking out, making noises, being out of one's seat, swearing, and name calling). Daily baseline observations of disruptive behavior were made in class. The intervention was a treatment package of several components:

- Posting classroom rules on the student's desk (e.g., sit in your seat unless you have permission to leave, raise your hand for permission to speak)
- Giving special instructions to the teacher (e.g., using the word "please" before a request was made of the student, standing close to the student)
- Providing tangible reinforcers for compliance and good behavior (e.g., praise, stickers exchangeable for prizes) and mild punishment for disruptive behaviors (e.g., taking away a sticker)

Figure 8.7 shows that the program was introduced in a multiple-baseline design across the three students. Two other students in the same class, of same in age and ethnicity, and also with diagnoses of disruptive behavior disorders were assessed over the course of the study but never received the intervention.

In the final follow-up phase, the program was completely withdrawn. (Musser et al., 2001.)

As the figure shows, the intervention led to change for each of the three students at each point that the intervention was introduced and not before. The pattern strongly suggests that the intervention rather than any extraneous influences accounted for the change. This conclusion is further bolstered by the two control students who were observed over time in the same class. Essentially these students remained in the baseline phase over the course of the study and continued to perform at the same level over time.

The example shows the practical utility of the designs. One can intervene on a small scale (e.g., the first baseline) to see if the intervention is working or working sufficiently well. Then as change is evident, the intervention is extended to other baselines (individuals or behaviors). Of course, that an intervention works or does not work on one baseline does not necessarily mean the intervention will have the same effect on the others. However, if the first baseline shows little or no change once the intervention is introduced, it is better to go back to the drawing board and beef up the intervention rather than crossing one's fingers in hopes that "maybe the next behavior will be different."

8.5.3: Design Variations

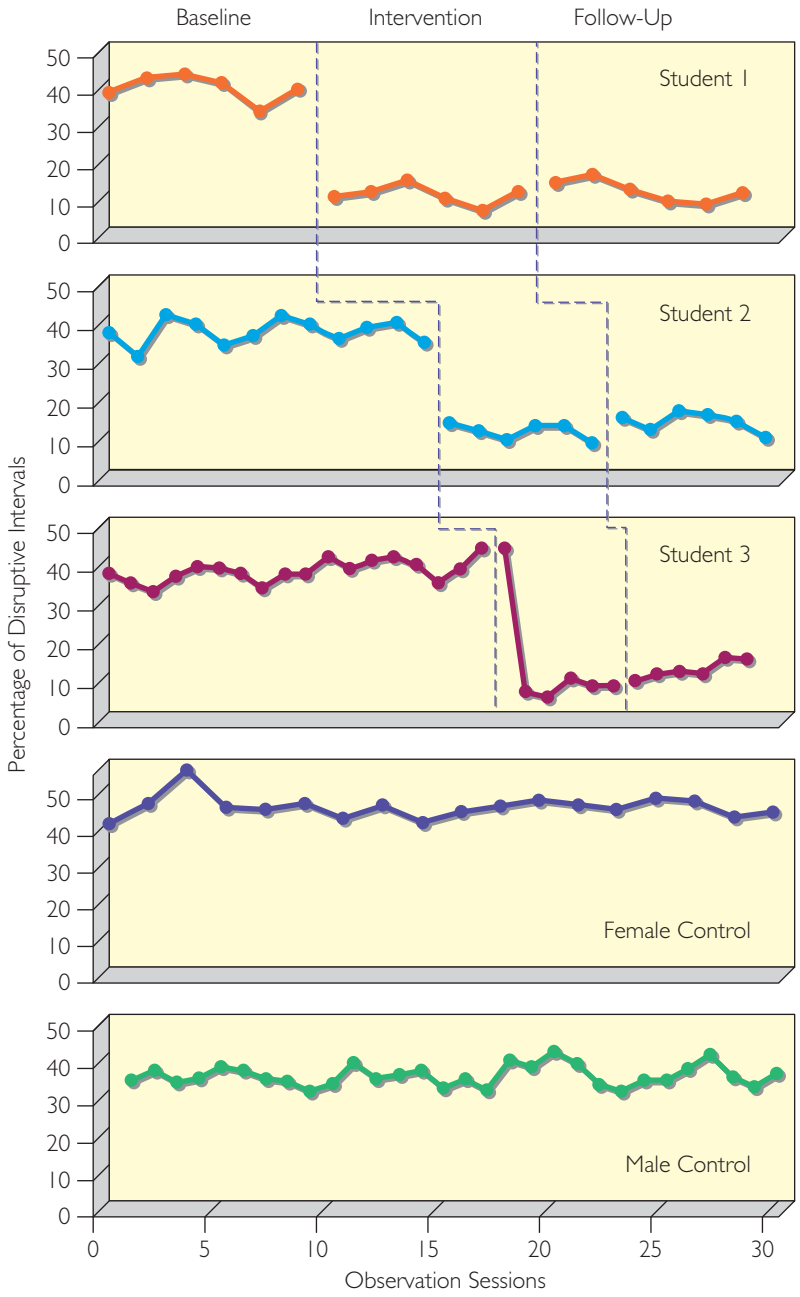
Multiple-baseline designs vary depending on whether the baselines refer to *different behaviors*, *individuals*, *situations*, *settings*, or *time periods*. I have already mentioned the version across individuals. A multiple-baseline across *different situations*, *settings*, or *time periods* of the day in which observations are obtained. This example focused on the safety of healthcare workers in the context of performing surgery (Cunningham & Austin, 2007). Healthcare workers suffer many injuries as a function of working with hazardous procedures or materials. Some states have enacted laws to help protect workers from "sharps injuries" (e.g., being stuck with a needle), given the special risk of such injuries for employees (e.g., HIV/AIDS). This study focused on the exchange of instruments between the surgeon and scrub nurse. The goal was to increase the use of the "hands-free technique" that requires that a neutral zone be established between the surgeon and nurse. This neutral zone is a place where the instruments are put as the instruments are exchanged. In this way, the two people do not touch the instrument at the same time and the risk of sharps injuries is greatly reduced.

This was a multiple-baseline design across settings: two settings were selected, namely:

- An operating room of an inpatient surgery unit
- An operating room of an outpatient surgery unit of a hospital serving a nine-county region in a Midwestern state in the United States

Figure 8.7: Disruptive Behavior (Percentage of Intervals) of Special Education Students

The intervention was introduced in a multiple-baseline design across three students. Two similar children (bottom two graphs) served as controls; their behavior was assessed over the course of the program but never received the intervention.



Observations were conducted during surgical procedures for 30 minutes, beginning at the time when opening incision was made in the patient. Observers were in the operating room, collected information, and recorded all exchanges as either hand-to-hand (unsafe) or neutral zone (safe, hands-free procedure). The percentage of these exchanges that were hands-free constituted the dependent measure. The intervention consisted of goal setting, task clarification, and feedback to use the safe-exchange procedure. At the beginning of the intervention phase, staff were informed of the hospital policy, which included use of hands-free procedure and set the goal to increase the percentage of hands-free exchanges. Hospital policy aimed at 75%, but the rate was only at 32%. Modeling was used to convey the exact ways of making the exchanges. Also, feedback was provided to staff regarding the weekly percentages and whether the goal was met. At these meetings, praise was provided for improvements in the percentages.

Figure 8.8 shows that the intervention was introduced in a multiple-baseline design across two surgery settings.

When the intervention was introduced to the inpatient operating room (top of figure), the percentage of safe exchanges increased sharply, so to speak. No changes were evident in the outpatient operating room where the intervention had yet to be introduced. When the intervention was introduced there, improvements were evident as well. There was only 1 day when the surgeon could not reach for the instrument in the neutral zone, as noted on the figure. Overall, the results convey that behavior changed when the intervention was introduced and not before. The design added a third phase in order to check to see if the behaviors were maintained. Approximately 5 months after the end of the intervention phase (the intervention had been suspended), both units were observed for a week. As evident in the figure, the effects of the intervention were maintained.

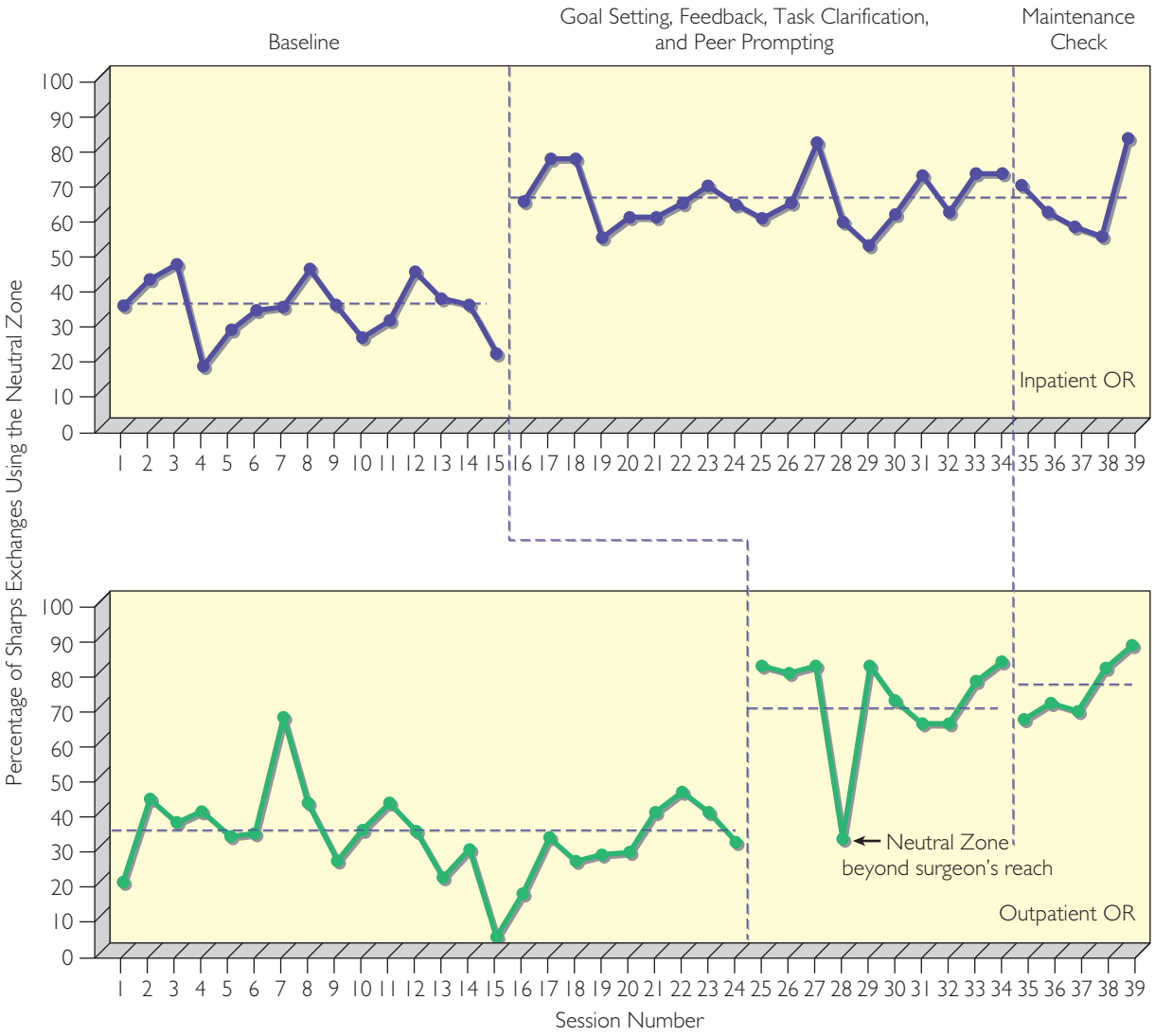
> When a change in behavior is required in two or more situations (e.g., home, school), the multiple-baseline design across situations or settings is especially useful. The intervention is first implemented in one situation and, if effective, is extended gradually to other situations.

The intervention is extended until all situations in which baseline data were gathered are included. As evident in the examples, the intervention is the most likely reason that explains the change. History, maturation, and other threats are not easily invoked to obtain the very special pattern of staggered changes across the multiple baselines.

Figure 8.8: Intervention Introduced in a Multiple-Baseline Design across Two Surgery Settings

Percentage of sharp instruments exchanged using the neutral zone (hands-free safe procedure) across inpatient and outpatient operating rooms. The solid lines in each phase represent the mean (average) for that phase.

(**Source:** Cunningham & Austin, 2007)



8.5.4: Considerations in Using the Designs

The multiple-baseline designs demonstrate the effect of the intervention without a return-to-baseline conditions and a temporary loss of some of the gains achieved. This immediately takes off the table the understandable concerns, both ethical and practical, that may emerge in an ABAB design. And like the ABAB designs, multiple-baseline designs can demonstrate a causal relation between the intervention and behavior change.

Two major considerations that affect the clarity of the demonstration are the number and the independence of the baselines:

1. The number of baselines adequate for a clear demonstration is difficult to specify. Two baselines are a bare minimum, but three or more strengthen the demonstration. With only two, a perfectly clear pattern in the data is essential to draw inferences about the impact of the intervention. More baselines (three and beyond) allow a little more room to see the onset of change when the intervention is introduced and allows for the possibility that one or more baselines may show the effect less clearly. The clarity of the demonstration across a given set of baselines is influenced by other factors such as the stability of the baseline data (e.g., few or no trends), the rapidity of behavior change after the intervention is implemented, and the magnitude of behavior change. Depending upon these factors, two or three baselines can provide a sufficiently convincing demonstration, as illustrated in the previous examples.
2. The design depends upon showing that the behavior changes when and only when the intervention is implemented. Ideally, behaviors still exposed to the baseline condition do not change until the intervention is applied. If they do, it suggests that maybe some factor other than the intervention may have led to the change. Occasionally, an intervention provided only for one or the first behavior may lead to changes in other behaviors that have yet to receive the intervention (e.g., Whalen, Schreibman, & Ingersoll, 2006). Some behaviors (e.g., communication, social interaction) may be pivotal to other activities and have ripple effects in changing other behaviors (e.g., Koegel & Kern-Koegel, 2006).

Similarly, in the multiple-baseline design *across individuals*, it is possible that altering the behavior of one person influences other persons who have yet to receive the intervention. In investigations in situations where one person can observe the performance of others, such as classmates at school or siblings at home, changes in the behavior of one person occasionally result in changes in other persons. For example, a program designed to reduce thumb sucking in a 9-year-old boy was very effective in eliminating the behavior (Watson, Meeks, Dufrene, & Lindsay, 2002). No intervention was provided to the boy's 5-year-old brother whose thumb sucking also was eliminated. It could have been that the brother who received the intervention was a cue for the desired behavior or modeled the behavior. The interpretation may not be clear in multiple-baseline designs when intervention effects spread in this way. Similarly, in the multiple-baseline design across *situations*, *settings*, or *time periods*, altering the behavior of the person in one situation may lead to carryover of performance across other situations. In this case too, changes before the intervention is applied in any multiple-baseline design can introduce ambiguity into the evaluation. In general, the spread of effects across different baselines before the intervention is introduced to each one appears to be the exception rather than the rule. When such generalized effects are present, features from other single-case designs (e.g., a brief reversal phase) can be added in separate experimental phases to demonstrate a causal relation between the intervention and behavior change.

> Multiple-baseline designs are user-friendly in education, business, and other settings in everyday life because the intervention is applied in a gradual or sequential fashion.

The investigator may wish to change many different behaviors of an individual (or different classrooms in a school, or in different schools). Rather than introducing the intervention to all of these at once, the program initially focuses on only one of these, which is often more feasible as a point of departure. In addition, if the intervention is effective, then it can be extended to all of the other behaviors for which change is desired. As importantly, if the intervention is not effective or not effective enough to achieve important changes, it can be altered or improved before it is extended. Thus, multiple-baseline designs have these additional practical advantages in implementing an intervention program.

8.6: Changing-Criterion Designs

8.6 Determine the functionality of the changing-criterion design

Changing-criterion designs demonstrate the effect of the intervention by showing that behavior matches a criterion for performance that is set for either reinforcement or punishment. As the criterion is repeatedly changed, behavior increases or decreases to match that criterion.

A causal relation between the intervention and behavior is demonstrated if the behavior matches the constantly changing criterion for performance.

8.6.1: Description

The changing-criterion design begins with a baseline phase in which ongoing observations of a single behavior are made for one or more persons. After the baseline (or A) phase, the intervention (or B) phase is begun. The unique feature of a changing-criterion design is the use of several *sub-phases* (b_1, b_2, to b_n). I refer to them as sub-phases (little b_n) because they are all in the intervention (B) phase; the number of these sub-phases can vary up to any number (n) within the intervention phase. During the intervention phase, a criterion is set for performance. For example, in programs based on the use of reinforcing consequences, the client is instructed that he or she will receive the consequences if a certain level of performance is achieved (e.g., completing three math problems from a baseline mean of 0). For each math session that performance meets or surpasses the criterion, the consequence is provided. As performance meets that criterion, the criterion is made slightly more stringent (e.g., six or seven math problems). This continues in a few sub-phases in which the criterion is repeatedly changed (e.g., up to a total of 10 problems that are assigned in each math class).

A more familiar illustration might be in the context of exercise. Baseline may reveal that the person never exercises (0 minutes per day). The intervention phase may begin by setting a criterion such as 10 minutes of exercise per day. If the criterion is met or exceeded (10 or more minutes), the client may earn a reinforcing consequence (e.g., special privilege at home, money toward purchasing a desired item). Whether the criterion is met is assessed each day. Only if performance meets or surpasses the criterion will the consequence be earned. If performance consistently meets the criterion for several days, the criterion is increased slightly (e.g., 20 minutes of exercise). As performance stabilizes at this new level, the criterion is again shifted upward to another level. The criterion continues to be altered in this manner until the desired level of performance (e.g., exercise) is met.

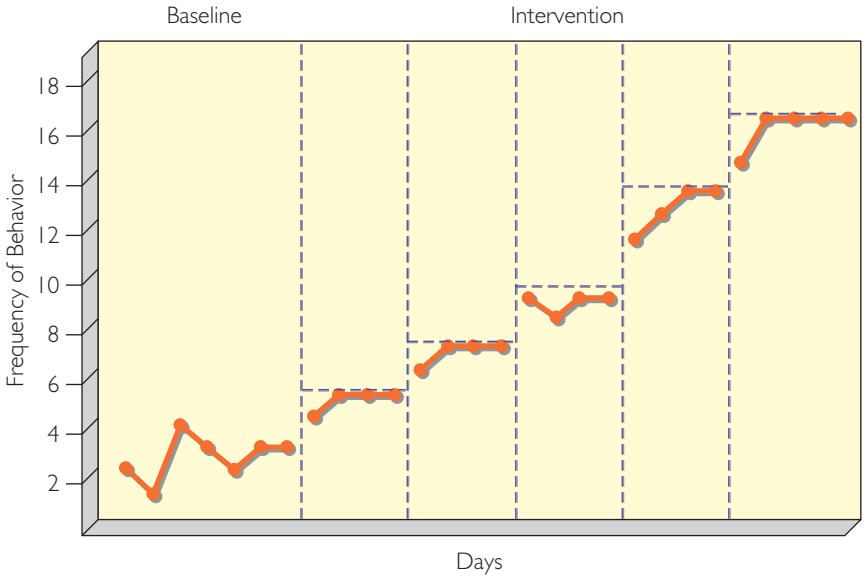
> Whether the criterion is consistently met does not necessarily require perfection; one looks at the pattern to see if performance jumps to the new criterion and hovers closely to that most of the time.

Figure 8.9 provides a hypothetical example of the changing-criterion design and shows a baseline phase that is followed by an intervention phase.

Within the intervention phase, several sub-phases are delineated (by vertical dashed lines). In each sub-phase, a different criterion for performance is specified (dashed horizontal line within each sub-phase). As performance stabilizes and consistently meets the criterion, the criterion is made more stringent. The criterion is changed repeatedly over the course of the design until the goal is achieved. The effect of the intervention is demonstrated if the behavior matches the criterion repeatedly as that criterion is changed. The logic of single-case designs is based on description, prediction, and testing of predictions in varied phases, as detailed in the discussion of the ABAB design. The logic still applies here with the mini-phases serving in the role of description and prediction.

Figure 8.9: Hypothetical Example of the Changing-Criterion Design

Hypothetical example of a changing-criterion design in which several sub-phases are presented during the intervention phase. The sub-phases differ in the criterion (dashed line) for performance that is required of the client.



8.6.2: Illustration

As an illustration, this study focused on a 15-year-old girl named Amy with insulin-dependent diabetes. She had been instructed to check her blood sugar 6–12 times per day (Allen & Evans, 2001). Among the challenges she faced was avoiding hypoglycemia (low blood sugar), which is extremely unpleasant and characterized by symptoms of:

- Dizziness
- Sweating
- Headaches
- Impaired vision

This can also lead to seizures and loss of consciousness. Children and their parents often are hyper vigilant to do anything to avoid low blood sugar, including deliberately maintaining high blood glucose levels. The result of maintaining high levels can be poor metabolic control and increased health risk for complications (e.g., blindness, renal failure, nerve damage, and heart disease). Amy was checking her blood glucose levels 80–90 times per day (cost about $600 per week) and was maintaining her blood glucose levels too high.

A blood glucose monitor was used that automatically recorded the number of checks (up to 100 checks) and then downloaded the information to a computer.

The test included:

- A finger prick
- Application of the blood to a reagent test strip
- Insertion of the strip into the monitor
- A display of glucose levels

An intervention was used to decrease the number of times blood glucose checks were made each day. Amy's parents gradually reduced access to the materials (test strips) that were needed for the test. A changing-criterion design was used in which fewer and fewer tests were allowed. If Amy met the criterion, she was allowed to earn a maximum of five additional tests (blood glucose checks). (Engaging in the tests was a highly preferred activity and was used as a reinforcing consequence; other consequences rewards could have been used.) Access to the test materials was reduced gradually over time. The parents selected the criterion of how many tests (test strips) would be available in each sub-phase. As shown in Figure 8.10, the criterion first dropped by 20 checks and then by smaller increments. Over a 9-month period, Amy decreased her use of monitoring from over 80 times per day to 12. Better metabolic control was also achieved; by the end of the 9 months, blood glucose levels were at or near the target levels (i.e., neither hypo- nor hyper-glucose levels).

One can see from the figure that the responses (number of checks) followed in a step-like fashion as the criterion changed. This is obvious when the first criterion was used after baseline ended (criterion of 60) and then in the large step after that (down to 40) and then to the next step (down to 20). It is very plausible that the intervention was responsible for change. Other influences (e.g., various threats to internal validity) would not be very plausible to explain the step-like changes that matched a changing criterion.

8.6.3: Design Variations

The usual version of the design consists of changing the criteria so that more and more or better performance is required to earn the consequences. One looks for directional change, i.e., progress in one direction toward improved behavior. A variation sometimes used is one in which a brief period is implemented during the intervention in which the criterion is temporarily made *less* stringent. That is, the individual performs better and better and matches the criteria and then a slight lowering of the criterion is implemented.

> One implements a phase in which the criterion is altered slightly so that there are bidirectional changes (improvements and decrements) in behavior.

This is not a complete return-to-baseline as in an ABAB design, but rather a slight change in the criterion to make it less stringent. Consider the sub-phase in which a less stringent criterion is used as sort of a "mini-reversal" phase. This is still the intervention phase, but the criterion is altered so that the expected change in behavior is opposite from the changes in the previous sub-phase.

An example is provided from an intervention with an 11-year-old boy named George with separation anxiety disorder, a psychiatric disorder in which the child is very extremely upset by separating from a parent or caregiver (Flood & Wilder, 2004). Difficulties in separating from parents at a young age are common and part of normal development. For some children, this may continue beyond early childhood and reflect more severe reactions that impair their daily functioning. George had intense emotional reactions and could not allow his mother to leave without displaying them.

The intervention was provided on an outpatient basis twice per week. Each of the sessions lasted up to 90 minutes. The intervention consisted of providing reinforcers for the absence of emotional behaviors and increases in the amount of time George could separate from his mother without these reactions. During baseline, George and his mother were in the treatment room, and the mother attempted to leave by saying she had something to do and would be back soon. Because George showed strong emotional reactions, she stayed. During the intervention sessions, the mother began in the room but left for varying periods. A time was selected, in discussion with George, about how much time he could remain apart from his mother. If George met this time and did not cry, whine, or

Figure 8.10: Number of Blood Glucose Monitoring Checks Conducted During Last 10 Days at Each Criterion Level

Maximum test strips allotted at each level are indicated by dashed lines (the changing criteria) and corresponding numbers of checks. The number of checks above the criterion level reflects the number of additional test strips earned by Amy.

(**Source:** Allen & Evans, 2001)

show other emotional behavior, he could have access to various toys and games for 30 minutes or could receive a small piece of candy or a gift certificate that could be exchanged at a local toy store. If he did not meet the time, he would have a chance in the next session. While the mother was away (outside of the room or later off the premises), she would be called back if George had an emotional reaction to the separation. That ended the session.

More and more minutes free from emotional reactions were required to earn the reinforcer. Although the demonstration seemed clear—in fact the criterion was matched for all but 1 day (day 30), a mini-reversal was introduced by decreasing the requirement to earn the reinforcer from 24 to 18 minutes (see sessions 19 and 20 in the figure). That is, less behavior was required of George than in the previous sub-phase. Behavior declined to the new criterion. Then, the criteria were made more stringent. Finally, in the last phase of the study, the criterion was lowered (made less stringent) on four occasions and behavior fell to that level too. Throughout the study, performance matched the criterion. The demonstration is particularly strong by showing changes in both directions, i.e., bidirectional changes, as a function of the changing criteria.

In this example, there was little ambiguity about the effect of the intervention. In changing-criterion designs where behavior does not show this close correspondence between behavior and the shifting criteria, a bidirectional change may be particularly useful.

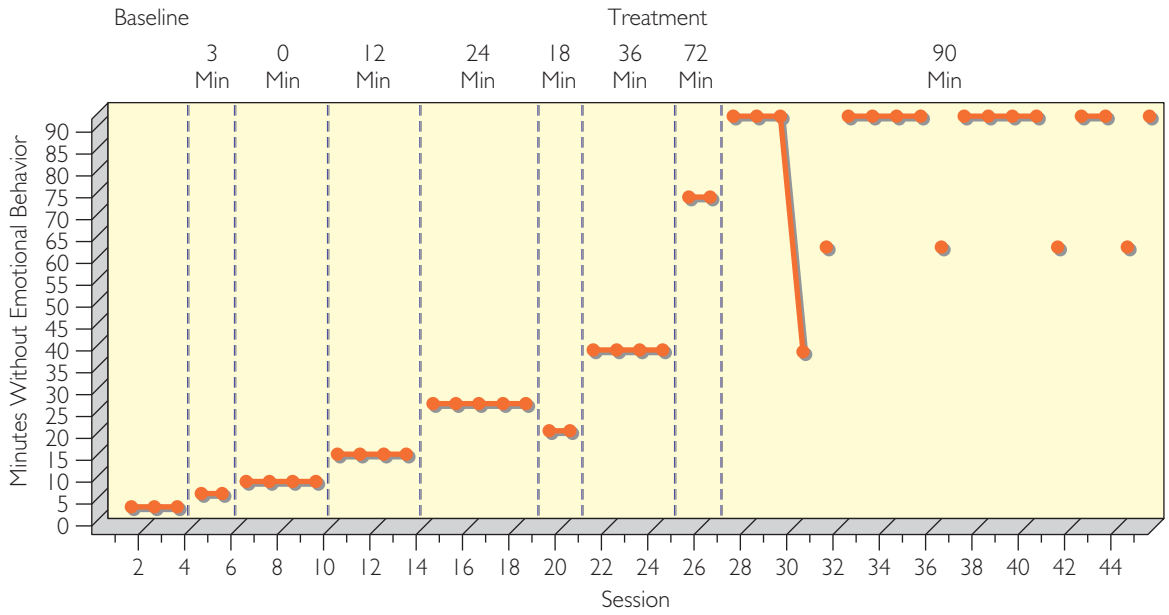
> When performance does not closely correspond to the criteria, the influence of the intervention may be difficult to detect. Adding a phase in which behavior changes in the opposite direction to follow a criterion reduces the ambiguity about the influence of the intervention.

Bidirectional changes are much less plausibly explained by extraneous factors unrelated to the intervention than are unidirectional changes.

Figure 8.11: A Baseline Phase and the Intervention Sub-Phases

Minutes without emotional behavior while George's mother is out of the room. Solid lines in the data represent jointly established therapist and participant goals.

(**Source:** Flood & Wilder, 2004)



The most common use of the changing-criterion design is the one in which criteria are altered repeatedly for improved performance (i.e., no mini-reversal). The design is flexible so that the number of changes in criteria and how large those criterion shifts are can vary as a function of how well, poorly, or consistently the client is performing. The critical feature of the design is trying to demonstrate that a change in the criterion during the intervention phase is matched or approximated by shifts in the client's performance in response to these changes.

8.6.4: Considerations in Using the Designs

The design depends upon repeatedly changing the performance criterion and examining behavior relative to the new criterion. *The design is especially well suited to those terminal responses that are arrived at or approximated gradually.* In so many areas of life (e.g., developing a skill; improving along some dimension such as strength, duration of an activity, accuracy; developing or eliminating habits), the goals are approached gradually rather than all at once so the changing-criterion design is quite useful. Similarly, many educational applications focus on gradual development of skills (e.g., mastering math problems, reading more complex materials, amount of time exercising, or practicing music). Shaping these behaviors is consistent with gradually increasing a criterion for performance. Consequently, the design is very well suited to many applications in applied settings where progress is likely to be gradual.

Sometimes behavior changes may take large leaps. For example, the program may require the person to decrease cigarette smoking from a baseline rate of 30 per day to 25 (as the first criterion level for reinforcement). When the program is introduced, the person may go to 10 cigarettes per day and remain at that level or quit completely for reasons that are not understood. In general, if change occurs rapidly or in large steps and does not follow the gradual changes in the criterion, the specific effect of the intervention in altering behavior will not be clear. The changes may be influenced by some other factors (e.g., threats to internal validity). This is the reason why a mini-reversal phase (return to a prior criterion level but not back to baseline) is sometimes implemented, as noted previously. Showing that behavior changes in either direction (increase or decrease in performance) as the criterion is changed make a more powerful experimental demonstration. When there is a temporary lowering of the criterion, this is not a return to baseline and hence objections associated with reversal phases are less likely to apply.

> Overall, changing-criterion designs are quite useful. Changing a criterion gradually to achieve a terminal goal (e.g., improving the amount of homework completed, exercise, practice of some skill) can be very useful for developing the final goal behavior as well as for evaluating the impact of an intervention.

In general, the changing-criterion design is less persuasive in making threats to validity implausible than other single-case designs because the effects of extraneous events could account for a general increase or decrease in behavior.

The design depends upon showing a unidirectional change in behavior (increase or decrease) over time. However, extraneous events rather than the intervention could result in unidirectional changes. The demonstration is clear only if performance matches the criterion very closely and the criterion is changed several times. Making bi-directional changes in the criterion during the intervention phase strengthens the design, as mentioned previously.

8.7: Data Evaluation in Single-Case Research

8.7 Inspect how data is evaluated in single-case research

The ABAB, multiple-baseline, and changing-criterion designs are main variations of single-case designs, but there are many other options in use and combinations (Kazdin, 2011). The variations operate by the same logic of describing, predicting, and testing the predicted level of performance based on the collection of ongoing observations. Performance across different phases (rather than across different groups as the case in the more familiar between-group designs) serves as the basis of making comparisons. Designs, whether single-case or group, refer to the arrangements that allow us to draw valid inferences from the data and reduce the likelihood that threats to validity can explain the results. The arrangement is needed to draw inferences whether the intervention or manipulation was responsible for the change. This is quite separate from how the data themselves will be evaluated. Data evaluation has its unique and unfamiliar features in single-case methodology.

Data evaluation focuses on whether there was a change and whether that change is likely to be a reliable change rather than just due to fluctuations in the data. There would seem to be nothing to discuss here—almost all training in psychology, counseling, education, and indeed science more generally is based on statistical evaluation of the data. The primary, and almost exclusive, criterion is based on running one or more statistical tests. One enters the data on some spreadsheet or data entry program or imports from some automated data collection procedure all of the numbers into some software package for the appropriate statistical tests and finds out if the results are "significant," i.e., the conditions or groups meet conventional levels of statistical significance.

> Data in single-case research are evaluated with two different methods, non-statistical and statistical techniques. The primary and much more common method is non-statistical and is referred to as visual inspection.

Statistical tests are available for single-case designs, but they involve techniques that are somewhat less familiar (e.g., time-series analyses, randomization tests) and rarely covered in graduate training leading to research careers (see Kazdin, 2011). It is not the *availability* of statistical tests for the case that is the issue. Investigators working with single-case designs *prefer* non-statistical evaluation of the data. If you have this text for some class, it is likely you have had great exposure to between-group methods (e.g., from reading research) and statistical analyses. If that is the case, please fasten your mental seat belt for the methodological turbulence you may experience next.

8.8: Visual Inspection

8.8 Analyze how visual inspection is a non-statistical method of evaluating single-case research

> Non-statistical evaluation in single-case designs is referred to as visual inspection.
>
> *Visual inspection refers to reaching a judgment about the reliability or consistency of intervention effects across phases of the design by examining the graphed data.*
>
> There are many ways to graph data, but the usual way in which this is done is a simple line graph as evident in all of the examples in this chapter in which the data points are connected over time and within a given phase.

This allows one to see the pattern within a phase (in order to describe and predict) and to evaluate changes across the phases. Yet, there are quite specific criteria that are invoked to decide whether the changes are reliable once the data are graphed.

8.8.1: Criteria Used for Visual Inspection

Visual inspection primarily depends on four characteristics of the data that are related to the magnitude and the rate of the changes across phases (e.g., ABAB). These characteristics are based on evaluating a graph on which performance is plotted across phase in accord with the designs discussed previously in the chapter. The specific characteristics are listed and defined in Table 8.2 to provide a convenient summary.

The specific characteristics are:

1. See if there is a *change in means* (average scores) across *phases*. One looks for consistent changes in means across phases. A hypothetical example showing changes in means across phases is illustrated in an ABAB design in Figure 8.12 (top panel).
   Both intervention phases show an accelerating slope; the first and second baseline phases show no trend or a decelerating trend. The arrows point to the changes in level or the discontinuities associated with a change from one phase to another.
   As evident in the figure, performance on the average (horizontal dashed line in each phase) changed in

Table 8.2: Visual Inspection: Characteristics of the Data to Decide Whether Changes are Reliable

Characteristic	Definition
Changes in Means (averages)	The mean rate of the behavior changes from phase to phase in the expected direction.
Change in Trend Line	The direction of the trend line changes from phase to phase, as for example showing no trend or slope (horizontal line) in baseline and an accelerating trend during the intervention phase.
Shift in Level	When one phase changes to another, a level refers to the change in behavior from the last day of one phase (e.g., baseline) and the first day of the next phase (e.g., intervention). An abrupt shift facilitates data interpretation.
Latency of Change	The speed with which change occurs once the conditions (phases) are changed (e.g., baseline to intervention, intervention back to baseline).

NOTE: These criteria are invoked by examining the graphical display of the data.

response to the different baseline and intervention phases. Evidence of changes in means by itself may not be persuasive but contributes along with the other characteristics.

2. *Change in trend line.* As mentioned earlier, trend line refers to the tendency for the data to show a systematic increase or decrease over time. The alteration of phases within the design may show that the direction of behavior changes as the intervention is applied or withdrawn. Figure 8.12 (middle panel) illustrates a hypothetical example in which trends have changed over the course of the phase in an ABAB design. The initial baseline trend is reversed by the intervention, reinstated when the intervention is withdrawn, and again reversed in the final phase. A change in trend would still be an important criterion even if there were no accelerating or decelerating trend in baseline. A change from no trend (horizontal line) during baseline to a trend (increase or decrease in behavior) during the intervention phase would also suggest a reliable change.

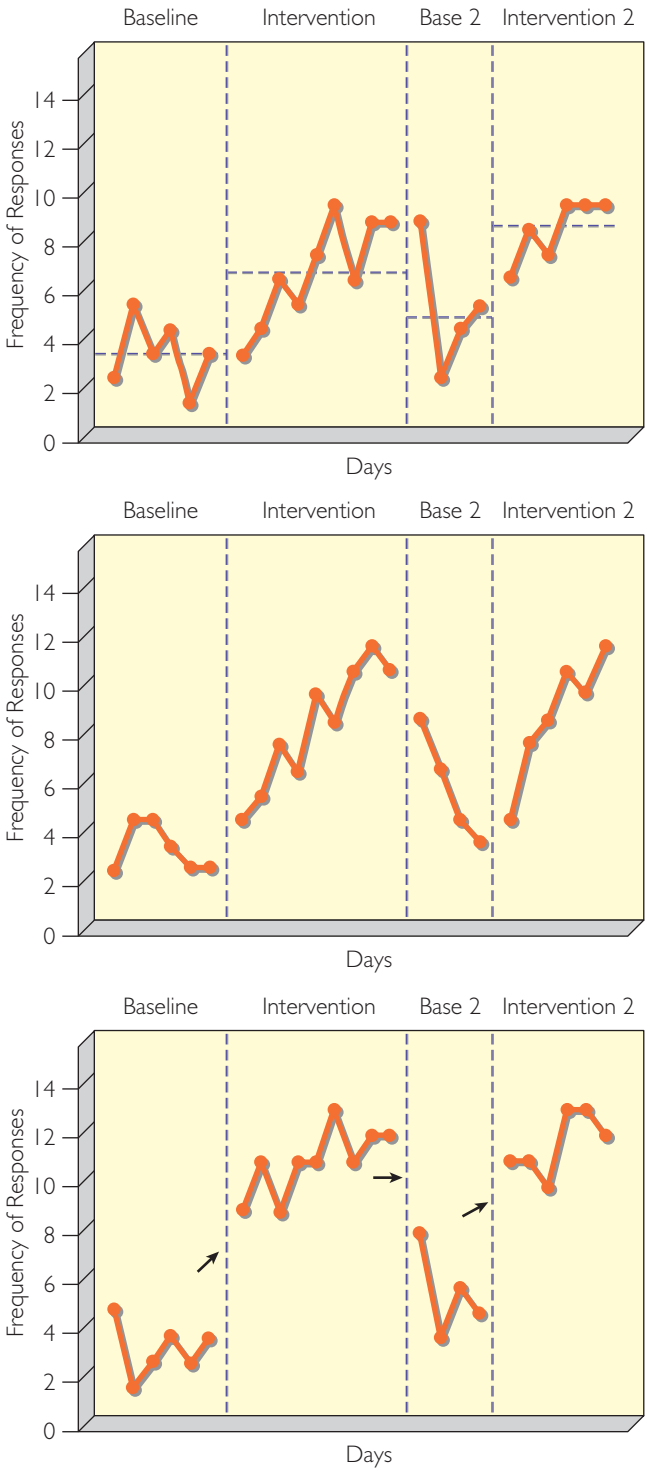
3. A *shift in level*, a little less familiar as a concept than are mean and trend.

   A shift in level refers to a break in the graphical display of the data or a discontinuity of performance from the end of one phase to the beginning of the next phase.

   A shift in level is independent of the change in mean. When one asks about what happened immediately after the intervention was implemented or withdrawn, the concern is over the level of performance. Figure 8.12 (bottom panel) shows change in level across phases in ABAB design. Whenever the phase was altered, behavior assumed a new rate, i.e., it shifted up or down rather quickly. It so happens that a change in level in this example would also be accompanied by a change in mean across the phases. However, level and mean changes do not necessarily go together. It is possible that a rapid change in level occurs but that the mean remains the same across phase or that the mean changes but no abrupt shift in level has occurred.

Figure 8.12: Data Evaluation

Top panel shows performance in an ABAB design in which there are clear changes in means (dashed lines) across phases. Middle panel shows changes in slope or trend from one phase to the next Bottom panel shows a shift in level.



4. The latency of the change that occurs when phases are altered is an important characteristic of the data for invoking visual inspection. Latency refers to the period between the onset or termination of one condition (e.g., intervention, return to baseline) and changes in performance. The more closely in time that the change occurs after a particular condition has been altered, the clearer the effect. There is a commonsense feature of this. If I tell my 10-year-old child to clean her room and she does this immediately (short or no latency), the chances are my request was the intervention responsible for change. If I tell that same child to clean her room and she does this 1 month later or right before getting dressed for her high school prom, my request could have been responsible but the long delay very much suggests that something else (e.g., packing her things to move to college) was involved.

8.8.2: Additional Information on Criteria Used for Visual Inspection

To convey what different latencies look like, consider the hypothetical data in Figure 8.13, which shows only the first two phases of separate ABAB designs.

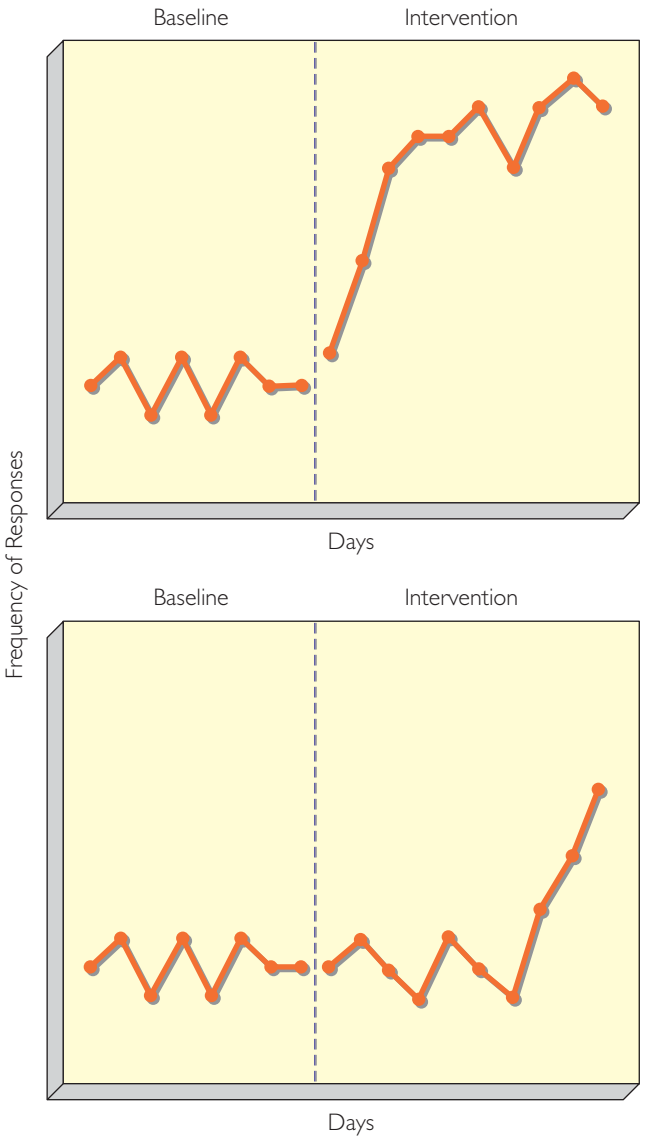
The changes in both top and bottom panels are reasonably clear. Yet as a general rule, as the latency between the onset of the intervention and behavior change increases, questions are more likely to arise about whether the intervention or extraneous factors accounted for change.

In the top panel, implementation of the intervention after baseline was associated with a rapid change in performance. In the bottom panel, the intervention did not immediately lead to change. The time between the onset of the intervention and behavior change was longer than in the top panel, and it is slightly less clear that the intervention may have led to the change. As a general rule, the shorter the period between the onset of the intervention and behavior change, the easier it is to infer that the intervention led to change.

Latency as a criterion for visual inspection cannot always be invoked to evaluate the impact of an intervention depending on the type of intervention and domain of functioning. For example, one would not expect rapid changes in applying a diet or exercise regimen to treat obesity. Weight reduction usually reflects gradual changes after interventions begin. If one plotted calories or minutes of exercise, one might look for a short latency, but not if one plotted weight loss. In contrast, stimulant medication is the primary treatment used to control hyperactivity among children diagnosed with attention deficit hyperactivity disorder. The medication usually produces rapid effects, and one can see changes on the day the medication is provided (one often sees a return to baseline levels on the same day as the stimulant is metabolized). More generally, drawing inferences about the intervention also includes considerations about how the intervention is likely to work (e.g., rapidly, gradually) and how that expectation fits the data pattern.

Figure 8.13: Hypothetical Examples of First AB Phases as Part of Larger ABAB Designs

Top panel shows that when the intervention was introduced, behavior changed rapidly. Bottom panel shows that when the intervention was introduced, behavior change was delayed.



To invoke visual inspection, one considers changes in means, trends, and levels and latency of change across phases. Often two or more of these characteristics go together. Yet they are separate characteristics of the data and can occur alone or in different combinations.

> Data evaluation and drawing inferences about the impact of the intervention require judging the extent to which these characteristics are evident across phases and whether the changes are consistent with the requirements of the particular design.

The individual components are important but one looks at the *gestalt* too, i.e., the parts all together and the whole they provide across all of the phases. When changes in mean, slope, and level are evident and latency of change is short, conclusions about the impact of the intervention are compelling.

In especially clear instances, the data across phases may not overlap.

> *Non-overlapping data refer to the pattern in which the values of the data points during the baseline phase do not approach any of the values of the data points attained during the intervention phase.*

For example, if one looks at the bottom panel of Figure 8.13, not one data point in baseline (A) was the same as or within the range of data points during the intervention (B). Non-overlapping data where little variability is evident, i.e., in real data, are even more impressive. In short, if there are changes in the means, levels, and trends, and short latencies across phases and the data do not overlap, there is little quibble about whether the changes are reliable. And that conclusion, based on the data characteristics, is reached without statistical tests and *p* levels.

Rather than giving a further example to invoke these data evaluation characteristics, the reader is encouraged to apply the four visual inspection criteria to examples already given in this chapter. As one of these examples, consider the intervention mentioned designed to make surgery safer for doctors and nurses while they were exchanging surgical instruments (Cunningham & Austin, 2007). As noted in Figure 8.8, this was a multiple-baseline across two operating rooms. In the figure one can readily see changes in means and level from baseline to intervention phases and an immediate change (short latency) as soon as the intervention went into effect. Trend (flat line) in baseline and intervention phases did not change but that does not weaken the interpretation in any way. All but one data point were nonoverlapping in the study from baseline to intervention phases. From the example, what can we say? First, the multiple-baseline design criteria were met, namely, change occurred when and only when the intervention was introduced. Second, from visual inspection, we can conclude that the results are reliable and not very likely to be due to fluctuations in the data. The effects are strong and perhaps you as a reader would agree reliable. Worth adding, statistical evaluation of these data would be likely to show huge effects.

8.8.3: Considerations in Using Visual Inspection

Visual inspection has enjoyed widespread use in identifying effective interventions, and these effects have been replicated extensively. Basic experimental laboratory research with human and nonhuman animals (e.g., on learning, decision making, choice) and applied research (e.g., in education, rehabilitation, and psychological treatment) have relied on single-case designs and visual inspection as a method of data evaluation (Kazdin, 2013a; Madden, 2013).[3] In applied research, major advances in developing evidence-based interventions (e.g., treatment of drug addiction, autistic spectrum disorders, self-injury, and many more domains too numerous to list) have emerged using single-case designs. Thus, whatever initial reticence there is toward visual inspection must be tempered by now extensive literatures that have generated replicable scientific findings. Nevertheless, there are major concerns, and these especially emerge when the pattern of results is not as clear as many of my hypothetical and real examples illustrated. These concerns are:

1. It would seem that "visual inspection" is merely a fancy term for subjective judgment and therefore riddled with biases and personal preferences. Perhaps visual inspection, when I apply it to my data, shows great effects, but when I look at your data, the intervention effect is not so clear. After all, if data evaluation is based on visually examining the pattern of the data, intervention effects (like beauty) might be in the eyes of the beholder.[4] As I note later, statistical evaluation in traditional research designs has its own subjective judgment, but that is not a cogent reply to the concern. Yes, subjective judgment enters into visual inspection.
2. Decisions about the reliability of change through visual inspection require integrating many factors (changes in means, levels, and trends as well as the background variables, such as variability, stability, and replication of effects within or across subjects). There are no concrete decision rules to help determine whether a particular demonstration shows or fails to show a reliable effect. Judges, even when they are experts in single-case research, often disagree about particular data patterns and whether the effects were reliable (e.g., Normand & Bailey, 2006; Park, Marascuilo, & Gaylord-Ross, 1990; Stewart, Carr, Brandt, & McHenry, 2007). Perhaps as the strength of interventions becomes less clear, subjective judgment would play an increasingly greater role in how the effects are interpreted.
3. Human judges are subject to all sorts of influences that are below awareness, a fascinating area of research in its own right (e.g., Bargh & Morsella, 2008). It is often the case that we report what influenced (e.g., in finding another person attractive) but in fact the data show other influences outside of our awareness firmly guided our decision (e.g., Hill et al., 2013; Pazda, Elliot, & Greitemeyer, 2012). I mention this because judgments about the effects of an intervention via visual inspection are influenced by the extent to which the rater finds

the intervention acceptable, reasonable, and appropriate for the treatment goal (Spirrison & Mauney, 1994). More acceptable interventions are rated as more effective whether or not the data actually support that.

4. Visual inspection is not very useful in detecting small effects. A goal in devising the method was to emphasize strong intervention effects. Yet, with the perspective of time, we have learned that most interventions do not produce strong intervention effects (Parker, Cryer, & Byrns, 2006). So the rationale of using visual inspection as a filter to detect only strong effects is an ideal not routinely met. Also, we often want to detect small effects. These might be the basis of developing the intervention further or for applying a low-cost intervention to many people, some of whom might profit. Visual inspection is likely to miss such effects given the need for a stark data pattern to draw conclusions that the intervention was reliable.
5. Visual inspection cannot easily accommodate characteristics of the data that can obscure the detection of intervention effects. Continuous data collected for a given subject or group over time may have a characteristic referred to as *serial dependence*.[5] This refers to the possibility that data from one occasion to the next (Day 1, Day 3, etc.) from the ongoing observations over time may correlate with each other. Among the concerns, there is a hidden pattern or trend in the data points that is not detectable by merely looking at the graphed data. This means, the relations may not "pop out" so one can tell that little patterns within a phase or across larger time periods across phases are systematic but not recognizable. I mentioned accelerating or decelerating linear trends, but more subtle patterns (e.g., cycles) might be in the data as well but not "visible" on a graph. Not all data collected in single-case experiments have this serial dependence (one has to test for that with special statistics). But when the data do have this characteristic, agreement about the effects of the intervention is much less consistent across different judges who rely on visual inspection. If intervention effects are very strong, then visual inspection can more readily detect effects, but we already noted that often such strong effects are not evident.

On balance, what conclusions might be reasonable to make about visual inspection and to use as a data-evaluation method? Foremost among the conclusions is that single-case designs and visual inspection have been used effectively to demonstrate findings that are replicated and generalizable across samples.

The examples in this chapter (see all graphs that with real rather than hypothetical data) convey reliable effects that can be seen (without statistical evaluation). Those examples were taken from a vast literature, and I did not have to mine mountains of earth to find a few gold nuggets. Also, the specific criteria, when met (e.g., change in means, level, and so on), readily allow application of visual inspection. Often the criteria are not met or are incompletely met, and the utility and reliability of visual inspection are debatable. That is tantamount to situations in between-group research where the investigator notes (inappropriately) that a finding was statistically significant at $p < .10$ or $< .07$). (The technical term for these levels is "not significant!.") In other words, borderline, checkered, and not quite clear effects are a problem whether in single-case or between-group research.

That said, often there is disagreement among judges using visual inspection. That disagreement has been used as an argument to favor statistical analysis of the data as a supplement to or replacement of visual inspection. The attractive feature of statistical analysis is that once the statistic is decided, the result that is achieved usually is consistent across investigators. And the final result (statistical significance) is not altered by the judgment of the investigator. Yet, statistics are not the arbiter of what is a "true" effect, and there are scores of statistical options that do not always lead one to the same conclusions. But it is fair to say that statistical tests make some of the decision making more replicable by other investigators than would visual inspection.

8.9: Statistical Evaluation

8.9 Express the causes that make statistical evaluation an important tool for evaluating single-case research

Visual inspection constitutes the dominant method of data evaluation in single-case research. Interest in statistical analyses of single-case data emerged from several considerations:

1. We want a consistent way to identify if intervention effects are present. Statistical tests could provide a consistent way to do that and to circumvent the highly variable application of visual inspection.
2. Hidden features in the data (subtle trends that escape merely looking at the graphed data) provide artifacts that bias visual inspection—those biases can obscure real interventions effects as well as suggest effects where there are none. Statistical tests, or at least some, can take these into account and provide information about whether there is still a real (reliable) effect of the intervention.
3. The vast majority of journals in psychology, counseling, education, and for that matter in the social, biological, and natural sciences more generally rely on statistical evaluation. To those unfamiliar with the tradition of single-case methods are likely to view such

designs as not real or "rigorous" science. Including statistical analyses in a single-case project may allay concerns about that.

4. Getting one's research published is likely to be impeded by not using statistical analyses. There are journals I mentioned that publish single-case research but less often is such research included in mainstream journals in clinical psychology, psychiatry, counseling, and education, but there are exceptions.
5. There are many circumstances in which we may want to be able to detect small but reliable intervention effects that visual inspection may be ill suited for. Small but reliable changes may be very important given the significance of the focus, ease of delivery of the intervention, and the larger impact these changes have across many people.

For example, standard medical practice includes having the physicians say something to their patients who smoke cigarettes. Physician visits are relatively brief (median = 12–15 minutes) in the United States. During the visit, advice from the physician or nurse can have a small but reliable effect on smoking. The physician says something like the following to patients who are cigarette smokers: "I think it important for you to quit tobacco use now," or "as your clinician I want you to know that quitting tobacco is the most important thing you can do to protect your health." The comments lead to approximately 2.5% increment in abstinence rates of smoking compared to no intervention (e.g., Rice & Stead, 2008; Stead, Bergson, & Lancaster, 2008). The 2.5% is pretty puny when our goal is more like 100%. Yet, this is a low-cost, easily administered intervention and might save some lives. When we are dealing with lives, one is important, especially if it is ours. But the main point is that small effects and "weak" interventions can be of great value, and we want to be able to detect them perhaps because of their value on their own or as a basis for building on them to develop stronger interventions. Visual inspection may not detect small changes that are reliable, something we have learned from studies providing evidence for this very point.

> Statistical analyses may help determine whether the intervention had a reliable, even though undramatic, effect on behavior.

8.9.1: Statistical Tests

Single-case designs are rarely taught in undergraduate or graduate education. Understandably, the designs and their dominant method of data analysis (visual inspection) are not well known. The vast majority of investigators who use single-case designs do not use statistical test for data evaluation. Thus, an even more esoteric area of specialization is statistical evaluation of single-case designs. The reason has to do with special features of single-case data.

> Special features of single-case data including many observations collected over time and over changing phases of varied durations, data across different baselines (multiple-baseline designs), subphases of brief duration (changing-criterion designs) require special considerations not handled by the usual statistics one learns in undergraduate and graduate school (e.g., *t* and *F* tests, multiple regression).

I mentioned that data obtained from a given subject on multiple occasions over time may be correlated or dependent. For many of the more familiar tests have assumptions within them that the data are not correlated in this way. Thus, use of those tests leads to biased results (e.g., inflated or deflated *t* or *F* values) and cannot be easily interpreted.

Several statistical tests have been developed and evaluated in relation to single-case experimental designs. They are names we rarely hear, see, or read about (e.g., the C statistic, last treatment day technique, randomization tests, double bootstrap methods, split-middle technique, and there are many others; see Kazdin, 2011). Also, it is difficult to find many published studies using these statistical techniques in major (well-known) and minor (lesser known) journals or—for that matter—in obscure, unknown journals where my dissertation is being consider for publication—fingers crossed.

Many of the available statistical tests for single-case designs are still undergoing evaluation, new variations are emerging, and the utility of these tests is being explored (see Kazdin, 2011). In short, many of the tests are not ready for prime time because they are not well understood. *Time-series analyses* are a noteworthy exception because they have an extensive history and have been used in many disciplines (e.g., economics, business, criminality) where data are collected over time and where there is interest in evaluating trends, changes, and sometimes the effects of interventions (e.g., Borckardt et al., 2008; Box, Jenkins, & Reinsel, 1994; McCleary & McDowall, 2012). Also, statistical software packages (e.g., SPSS, SAS, Sysat, Statistica, Stata) include time-series analyses so that they are readily available and accessible. I use the plural for time series by referring to analyses because there are multiple variations that need not be addressed here. Enough to say this is no one single version.

> **TIME-SERIES ANALYSIS ILLUSTRATION.** Time-series analyses compare data over time for separate phases for an individual subject or group of subjects. The analyses examine whether there is a statistically significant change in level and trend from one phase to the next. Thus, the change is from phase A to B. The analyses can be applied to single-case designs in which there is a change in conditions across phases. For example, in ABAB designs, separate comparisons can be made for each set of adjacent phases (e.g., A_1B_1, A_2B_2, B_1A_2). In multiple-baseline designs, baseline (A) and intervention

(B) phases may be implemented across different responses, persons, or situations. Each baseline to intervention change (A to B) can be evaluated.

8.9.2: Additional Information on Statistical Tests

An example comes from a study that focused on the effectiveness of a cognitive-behavioral treatment (CBT) for insomnia among women treated for nonmetastatic breast cancer (Quesnel, Savard, Simard, Ivers, & Morin, 2003). Sleep disturbances are one of many psychological problems associated with the impact of cancer and characterize 30% to 50% of the patients. Patients participated if they completed radiation or chemotherapy and met diagnostic criteria for chronic insomnia disorder (by criteria of *the International Classification of Diseases* or the *Diagnostic and Statistical Manual of Mental Disorders*). Several measures were used involving multiple assessment methods, including clinical interviews, self-report daily diary and questionnaires, and electrophysiology (polysomnography) of sleep evaluated in a sleep lab. The intervention consisted of CBT conducted in eight weekly group sessions, approximately 90 minutes each. CBT included several components (stimulus control for insomnia, coping strategies, restructuring of dysfunctional thoughts). At pretreatment, posttreatment, and each follow-up assessment, an extensive battery of measures was completed. Electrophysiological measures of sleep were obtained at pretreatment, posttreatment, and the 6-month follow-up.

Figure 8.14 charts one of the continuous measures, which consists of a daily sleep diary kept by patients.

The measure was used to report several characteristics of sleep (e.g., use of alcohol or medication, bedtime hour, duration of awakenings, and others). As evident in the figure, CBT was introduced in a multiple-baseline design across participants. The results suggest through visual inspection that introduction of treatment was associated with decreases in total wake time, although the effects are less clear for participants 6 and 7. One can see that gains for those who responded to treatment appeared to be maintained at the follow-up periods.

A time-series analysis evaluated the statistical significance of the change for each participant across AB (baseline, treatment) phases. The analysis was selected because it takes into account otherwise difficult features to evaluate in the data through visual inspections (e.g., delays in the effect of an intervention, subtle cycles or patterns in the data, and serial dependence or the correlation of the data points over time because they are from the same subject), can detect reliable intervention effects even if the effects are small, and evaluates changes in level and slope.

> The analyses showed that all participants changed significantly either in level (participants 1, 2, 3, and 7) or trend line (subjects 4, 5, 6, 8) but no one changed in both. Thus, the statistical analyses convey that there was a reliable treatment effect; the complexity of the effect (level for some, trend for others) provides information that would be difficult to discern from visual inspection.

Several other analyses were completed (and not discussed here) that demonstrated reductions in depression and physical fatigue and improved cognitive functioning as well. At posttreatment and again at the 6-month follow-up, electrophysiological measures in the sleep lab revealed significant decreases in time awake and increases in sleep efficiency (proportion of time sleeping out of time in bed).

Time-series analysis was very helpful in evaluating data in which there was considerable variability for some of the participants in both baseline and intervention phases. Also, possible trends in the data and autocorrelation were modeled and handled by the analysis. By "modeled" I mean an algorithm is needed that best describes any pattern of data in baseline; this is required to determine whether intervention reflects a significant change over and above that pattern. This is more than visual inspection can accomplish. Any trends in baseline, whether or not they could be easily detected by visual inspection, were readily incorporated in evaluating changes from A to B phases. Perhaps one might argue that visual inspection would have been able to detect changes, perhaps for participants 2, 3, and 5 where the effects are among the clearest. Even here some statistic is needed to handle the invisible serial dependence and trends that are not simple ascending or descending straight lines.

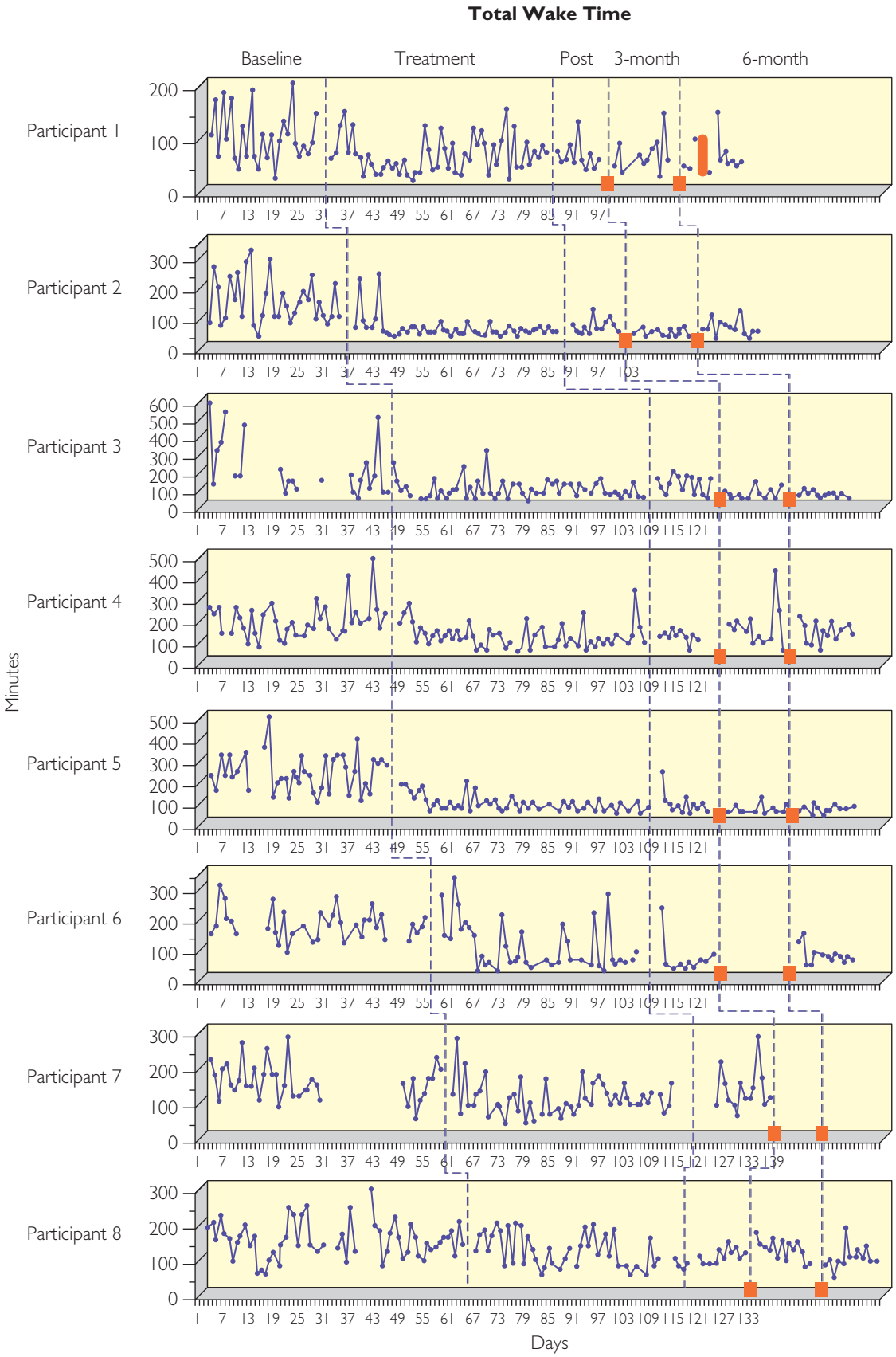
Time-series analyses are well developed, but they cannot always be easily applied to single-case data. To begin with, the design depends on having a sufficient number of data points. The data points are needed to determine the existence and pattern of serial dependence in the data and to derive the appropriate time-series analysis model for the data. The actual number of data points needed within each phase has been debated, and estimates have ranged from 20 to 100 (see Kazdin, 2011), with smaller than occasionally suggested (Borckardt et al., 2008). In many single-case designs, the number of data points within a phase is very small (5–10 in return to baseline phases or in the initial baseline phase of multiple-baseline design). Consequently too few data points may preclude the application of time-series analysis.

Second and related, time series is not a matter of plugging in numbers into a formula. There are steps performed on the data (by the computer program) that include model building, estimation, and evaluation (checking the model against the data). Within these steps are contained such tasks as how to best describe the pattern of autocorrelation; what estimates of parameters are needed to maximize the fit of the model to the data; and once estimated how the model has contained, addressed, or removed autocorrelation.

Figure 8.14: Daily Total Wake Time Obtained by Sleep Diaries for Each of Eight Participants Who Completed Treatment

Missing data (e.g., baseline, Participant 7) reflect the absence of the diary for those days. Treatment was cognitive-behavior therapy introduced in a multiple-baseline design across participants.

(**Source**: Quesnel et al., 2003)



Once these are complete, the analysis can test changes in level and slope associated with the intervention. Returning to a prior point, one reason many data points are needed for the analysis is to execute these initial steps to provide a good estimate of the model and parameters that fit the data. From this very cursory description, one can see that there is much to understand about time-series analyses. Although available software allows one to enter the data, it is important to understand the steps along the way to the final result and selection of the model. Misestimation of the model (characteristics) of the data and accepting or not accepting default options about the data within a program can lead to quite different effects (statistically significant or not statistically significant effects).

> Time-series analysis is especially useful when the criteria of visual inspection are not met. For example, when there is a trend toward improvement baseline, when variability is large, or when intervention effects are neither rapid nor marked, time-series analysis can detect intervention effects.

Also, the analysis is especially useful when the investigator is interested in drawing conclusions about changes in either level or trend. As reviewed previously, considerable data suggest that trend is not easily detected by visual inspection once one moves beyond simple ascending or descending straight lines. Time series represents a viable option, especially for large sets of data. The analyses require more sophistication than the more commonly used and familiar statistics in psychology, counseling, and education because there are multiple options in conducting time-series analyses and the options one selects and decisions about the data.

8.9.3: Considerations in Using Statistical Tests

It is important to mention statistical evaluation of single case primarily to note this is a possibility as an alternative to or supplement to visual inspection. Again worth noting is that statistical tests are not commonly used in single-case research, in part a matter of philosophy and history of the designs but also in light of the current status of the statistical tests:

1. Many statistical tests are available for single-case research, but there is no standard one or two of such tests that have been adopted for widespread use. Among the reasons, different statistical tests for the same data often yield different results (e.g., Lall & Levin, 2004; Manolov & Solanas, 2008; Parker & Brossart, 2003). Much of this has to do with how many phases in the design, duration of the phases, characteristics of the data (degree of autocorrelation), and assumptions underlying the tests and how these differentially influence a given statistical test (see Kazdin, 2011).
2. Some of the statistical tests that can be used (e.g., time-series, randomization tests) can compete with the demands of the experimental design. Using long phases to obtain data or making assignments of conditions (what is provided to the participant) based on needs of the statistic rather than the design introduces obstacles. Single-case designs make decisions about when to change phases based on patterns of the data within a phase, and some statistics (e.g., randomization tests) require other rules for deciding when the intervention is provided and for how long.
3. At present there are few resources to learn statistical tests for single-case designs, at least within the social sciences. This is understandable because the tests are not used very much largely because the vast majority of studies in psychology, counseling, and education do not collect ongoing data (many data points over time), which are fundamental to single-case research. Also, what tests should one learn, given that there is no commonly used or agreed-upon tests? And, different tests can produce quite different results, as I noted previously. A resource limitation is further reflected in the absence of most tests in statistical software packages. In contrast, for researchers evaluating data from between-group research, there is a wide selection of statistical packages (e.g., SPSS, SAS). These packages include multiple-statistical techniques for data evaluation, are constantly revised, and serve as one-stop shopping for many faculty members, postdocs, and graduate students doing empirical research. Software packages compete in their comprehensiveness of coverage and ease of use. In the case of single-case research, software is available to address specific tests (e.g., time-series analyses, randomization tests), but there is little with the range of coverage and ease of use that the more familiar statistical packages provide.
4. One rationale for considering statistical tests was the fact that visual inspection is not particularly good at detecting small effects, even if they are reliable. Perhaps, statistical tests could accomplish what the eyes cannot and show statistically significant effects in situations with fewer or none of the four visual inspection criteria are met. We do not really know how well statistics can detect differences in single-case research in these situations. At present, a few studies have shown that some statistical tests for single-cases designs are not very useful in detecting differences across phases unless the effects are very strong (e.g., effect sizes are very large –> 2.0) (e.g., Ferron & Sentovich, 2002; Manolov & Solanas, 2009).[6] Yet, when effects are very strong, visual inspection does quite well too.

There is a role for statistical evaluation in ways that could contribute greatly to single-case evaluation. Statistical significance testing is only one way of evaluating the

data. From the very inception of the development of tests of statistical significance in between-group research, there has been concern about the limitations. Statistical significance is dependent on sample size, gives a binary decision, and does not say anything about the magnitude or strength of the effect. One can readily conclude that there is no effect (not statistically significant) when in fact there is (*called Type II error*). Spanning decades but exerting influence more recently has been the view that statistical significance should be supplemented by, if not replaced with, some measure of the magnitude of effect. How large an effect is can be distinguished from whether the effect was statistically significant. Effect size has been the measure of magnitude of effect frequently advocated and does not suffer the same problems as does statistical significance.

I mention this because a related development is the use of meta-analysis as a way of reviewing and integrating empirical studies on a given topic by translating the results of these studies (e.g., changes on outcome measures, differences among groups) to a common metric (effect size). This allows the reviewer (meta-analyst) to draw conclusions about the findings in a given area and to quantify the strength of effects. In addition, one can ask questions about the data from many studies combined that were not addressed in any of the individual studies included in the meta-analysis. Thus novel findings can emerge from a meta-analysis. Between-group researchers engaged in intervention research are encouraged or required to provide effect size information, depending on the journal and discipline. Even when researchers do not provide that information, often effect sizes can be obtained from other statistics that are in the original article (e.g., means, standard deviations for various measures).

Contrast the situation with single-case designs and the use of visual inspection. Visual inspection from one study to the next does not provide a systematic way of integrating and combining many studies or of asking new questions based on a large integrated database. Without some formal, replicable way of combining studies, much of the single-case work is neglected or viewed as difficult to integrate. Over the years, many researchers have proposed effect size measures for single-case designs. In fact, over 40 different approaches for measure effect size have been proposed for single-case research (Swaminathan et al., 2008). None has been widely adopted and only recently have some of the alternatives been carefully evaluated and compared (e.g., Manolov & Solanas, 2008; Parker & Brossart, 2003; Parker & Hagan-Burke, 2007b). In short, there is no recommended method of computing effect size in single-case designs that is readily available and ready for prime time. The absence of a clear and widely used way of computing effect size limits the ability to accumulate and combine findings from single-case studies and integrating findings of single-case and between-group studies. This too is quite relevant background of the current interest in using statistical tests.

More and more studies use statistical tests to analyze single-case data, sometimes along with visual inspection (e.g., Bradshaw, 2003; Feather & Ronan, 2006; Levesque, Savard, Simard, Gauthier, & Ivers, 2004; Quesnel et al., 2003). Also, many articles have emerged that present new statistical tests for the single case, reanalyze prior data from published studies, or present new data to illustrate the analyses. Some of these articles compare multiple single-case statistical tests (e.g., Brossart, Parker, Olson, & Mahadevan, 2006; Parker & Brossart, 2003; Parker & Hagan-Burke, 2007a, b). While it remains to be the case that visual inspection dominates, statistical evaluation has been on the march.

> Two summary points ought to be emphasized in relation to the use of statistical tests. First, such tests represent an alternative to or a complementary method of evaluating the results of a single-case experiment. Second, statistical evaluation can permit accumulation of knowledge from many different investigations, even if they do not all use the same statistical tests.

Enormous gains have been made in between-group research by looking at large literatures and drawing quantitatively based conclusions. Combining studies that use visual inspection to reach conclusions and pose and answer new questions from such a data set have yet to emerge in single-case research. Findings from visual inspection risk continued neglect from a broad scientific community if they cannot be integrated in a way that effect size has permitted in between-group research. The solution is not merely applying currently used effect size estimates and applying them to single-case research. Characteristics of single-case assessments and data (e.g., ongoing assessments, influence of the number of data points, serial dependence) make the usual formula not directly applicable (see Shadish, Rindskopf, & Hedges, 2008).

On balance, visual inspection remains the standard way of evaluating data in single-case designs. This is preferred by investigators who conduct this research and in the journals that publish the research. When single-case designs were coming into their own, it made sense perhaps to carve out an identity by showing how visual inspection was unique and accomplished things that were not achieved by statistical tests. Visual inspection and statistical evaluation, very much like single-case designs and traditional between-group designs, are tools for drawing inferences. There is no need to limit one's tools, and indeed there are several disadvantages in doing so. Few studies provide statistical evaluation and evaluation of the individual data via visual inspection. When they do, they convey the critical point: the methods have a slightly different yield and each provides information the other did not provide (e.g., Brossart, Meythaler, Parker, McNamara, & Elliott, 2008; Feather & Ronan, 2006).

8.10: Evaluation of Single-Case Designs

8.10 Scrutinize the strengths and the weaknesses of single-case designs

Single-case designs have provided a viable methodology that has been used in thousands of studies, for a period spanning several decades, and have elaborated the nature of key psychological processes in basic research with diverse nonhuman animals (e.g., birds, rodents, primates, and so many others) as well as human populations. In relation to applied work in clinical psychology, counseling, education, and rehabilitation and community work, the approach has been used with age groups (e.g., from infants through the elderly), clinical populations (e.g., children with autism spectrum disorder, conduct disorder, anxiety; adults with schizophrenia, substance abuse disorders), and settings (e.g., all levels of educational settings, the home, prisons, military, college dorms and athletics, conversation and energy consumption in community samples—it is endless). Evidence-based interventions have emerged from the vast empirical literature using these designs.

8.10.1: Special Strengths and Contributions

Single-case designs have special strengths and contributions worth noting.

1. Strength 1 of Single-Case Designs: Expand the range of options and opportunities to evaluate intervention programs in everyday life as conducted in relation to diverse goals, domains, and settings (e.g., related to health, education, safety, conservation, and more).
2. Strengths 2 and 3 of Single-Case Designs: Provide a way to evaluate change and impact of interventions for a single case (e.g., particular person, group in a particular setting) without requiring the accumulation of many participants and assignment of these participants to various control or comparison groups.
3. Strengths 2 and 3 of Single-Case Designs: Provide ongoing feedback from the data to permit informed decision making that can help clients while the intervention is still in effect. Changes can be made based on the feedback as needed to improve client performance.
4. Strengths 4 and 5 of Single-Case Designs: Allow for the gradual or small-scale implementation of the intervention (e.g., across one individual or one behavior or one setting) before larger scale application. This allows one to try out, perfect, and modify the intervention as needed before the extension is made to the group as a whole or to other individuals.
5. Strengths 4 and 5 of Single-Case Designs: Permit investigation of rare problems among individuals who are not likely to be studied in between-group research because there is not a feasible way to recruit multiple participants with similar problems.

Each of the above strengths is discussed in detail in the following sections.

8.10.2: Strength 1 of Single-Case Designs

> First, they expand the range of options and opportunities to evaluate intervention programs in applied settings.

The world is filled with programs and interventions designed to help people (e.g., in day-care centers, special education and regular classrooms, everyday life such as fostering vaccinations, nutrition, exercise, promoting energy conservation) or increase skill performance (e.g., in sports, academic and athletic abilities), improve safety (e.g., in hospitals, in business), compliance with the law (not texting while driving, not speeding), and so on. The vast majority of such programs are not evaluated at all.

Do they help? Do they harm? Or, do they do neither but drain resources under the guise of helping?

What do you think?

Evaluation of effectiveness with between-group designs sometimes is an option, especially quasi-experimental designs where random assignment may not be possible (e.g., schools). Even in that context, meeting some of the demands of group research (e.g., obtaining control and comparison groups, ensuring a sufficient sample size to detect a difference if there is one [statistical power]) can preclude evaluation. Single-case methods provide a viable set of options to evaluate individual programs whether they are implemented on a small (one child, one classroom, one school) or larger scale.

8.10.3: Strengths 2 and 3 of Single-Case Designs

> Second, single-case designs provide a way to evaluate change and impact of interventions on a particular person or single setting (e.g., classroom, office, business).

For example, we learn from between-group research that a particular intervention is effective or among the best options for a particular problem we care about in relatives, friends, or ourselves (e.g., obesity, diabetes, blood pressure, or hair loss). Yet, we also need to know if the intervention is effective for a given individual undergoing care now. This is important because in applied work, we want

to make a difference in someone's life, to know if we have done that, and to know if the intervention was likely to be responsible for the change. Single-case designs are not only compatible with actually helping people, they also address the danger of believing one is making a difference without actually evaluating to see if one has.

> Third, assessment of single-case designs can provide ongoing feedback from the data and permit informed decision making to help clients while the intervention is still in effect.

In between-group research, the intervention is preplanned and administered in keeping with that plan. The impact of the intervention is evaluated at the end when the full intervention has been delivered (posttest assessment). Now that the data are in, one can evaluate the intervention. This makes sense for research (evaluating an important questions about a possibly effective intervention) but less so for helping specific individuals who receive the intervention (the questions individuals care about understandably are about themselves too).

Single-case designs allow for evaluation of impact *while* the intervention is in place. We can evaluate whether the intervention is achieving change and whether the change is at the level we desire or need. Decisions can be made *during* the intervention to improve outcome if the changes are not occurring or occurring in the desired degree.

Ongoing assessment during the intervention phase makes the designs quite user-friendly to the investigator (teacher, doctor, or other person responsible for the intervention) and the client (person or group intended to benefit).

8.10.4: Strengths 4 and 5 of Single-Case Designs

> Fourth, single-case designs allow for the gradual or small-scale implementation of the intervention (e.g., across one individual or one behavior or one setting).

With one or a few cases, one can implement the intervention and see in a preliminary way whether this is having an effect. This allows the investigator to modify the intervention on a small scale if needed before applying the intervention to the entire class, school, or other larger scale setting. If there is a strong intervention effect in the small-scale application with one or a few subjects, this does not necessarily mean that the effect will extend across all subjects or baselines. But the point here is that first starting out on a modest scale, across one phase for one or two individuals (ABAB) or across one baseline (in a multiple-baseline across individuals, situations, responses) helps the investigator preview the impact of the intervention as well as master implementation and some of the practical issues that may relate to its effectiveness.

> Finally, single-case designs permit investigation of rare problems among individuals who are not likely to be studied in between-group research.

There are many clinical problems that are relatively rare so that it would be extraordinarily difficult to recruit subjects for a large-scale treatment evaluation project. Occasionally, it is not so much that a problem is rare, but it comes to us (e.g., in clinical work, special education) embedded in other conditions (two or more additional psychiatric or physical disorders are present) or circumstances (e.g., special living conditions, parents who have limitations) and there would be no way to do a "group" study to understand or help develop and test an intervention. Yet single-case designs allow careful investigation of an individual client in such contexts and in which between-group research is not an option. This allows developing effective interventions for that individual.

8.10.5: Issues and Concerns

For individuals who have not used single-case designs, perhaps the greatest concern voiced about single-case research is external validity, and within this specifically the generality of the findings to other subjects than the one or a few included in the study. That is, the finding with one or a few subjects may not generalize to a lot of other people. Surely between-group research is better for generalizing the findings. Not really or at least not necessarily.

There are three main points to note about generality of findings in single-case as opposed to between-group experimental research:

1. Generality of findings has not been a problem for single-case research. That has to do with the type of interventions that usually are evaluated and their underpinnings (Kazdin, 2013a). Single-case research has been associated with interventions developed extensively from human and nonhuman animal research (e.g., principles of operant conditioning), and these interventions have wide generality, often across many species but more pertinent perhaps across many human populations (e.g., from infants to the elderly). Thus, the findings from single-case research have been widely applicable, not so much related to the designs as to the interventions the designs are used to evaluate.
2. Generality of findings from between-group research is not what we think and is hardly clear or automatic. In between-group research, participants are assigned to different conditions or groups. The differences between these groups on some measure(s) are evaluated statistically. Usually, between-group research focuses on means and measures of variability (e.g., standard deviation) and of course these are descriptive statistics about the groups in the study. We may know from

the means that the group that received some intervention is better overall ($p < .01$) than another group that received some other condition (no intervention). That usually reflects that the group means are different. But what is the generalizability of this finding? Does the finding (more effective) characterize and generalize to any particular individual or to most participants in that study? That is, we are not yet talking about generality of findings to other people not in the study. We have a prior question.

We usually cannot tell the extent to which individuals within a group study reflect the overall pattern (e.g., showed improvement) or improved in a way that made a difference. Group data are not analyzed in a way to see how many individuals actually responded well. The data are available within the study but rarely examined.

Thus, the generalizability of findings within a group study is not so clear. And just because lots of subjects were used, we cannot extrapolate that percentage (not provided) that might be a guide as to how many others in the population might respond.

3. Generalizability of a finding from a sample to a group depends on sampling carefully from the overall group. That is, generality requires some initial assurances that the sample (in my experiments) represents the group (e.g., all people? all college students? all ethnic and cultural groups?). Experiments in psychology, counseling, education, and other such areas rarely use random selection from a population—that would allow better generalizability from the sample to the population. Rather, experiments use random assignment to conditions and that is not especially relevant to generalization.

Single-case research allows one to see whether the findings are similar (generalize) among the subjects included in the design, when more than one subject is used. Both between-group and single-case designs have challenges in relation to generality. And in both cases, the key is replication of findings with a new set of participants.

Although generality is not an inherent problem in single-case research, there is a place where between-group handles the issue of generality of findings among subjects much better. Between-group research often studies moderators, i.e., those variables that may influence the direction or magnitude of change. The study of moderators can also be framed as a study of generality of effects, because the questions is "Does the intervention (experimental manipulation) effect generalize across subjects who show both (or all) levels of the moderator?"

For example, one might evaluate treatment (e.g., for depression) and propose that the treatment will be more effective with individuals who have no family history (in parents and siblings) of depression. We recruit individuals who meet some cutoff on measures of depression and who vary in having no history of depression in their family versus those who have at least one parent with a history of depression. We are posing the effectiveness of our treatment is influenced (moderated) by family history. Those with a family history may have stronger genetic loading for depression and stronger environmental loading (e.g., child rearing changes when one's parent is depressed). There are many ways to test this statistically, and there is no need to take the example to that point. One can see that we need groups for this study, a sample size to allow the comparisons (statistical power), and so on. If the moderator in fact is related to the results, this means that the intervention is differentially effective with one group more than another or that the intervention works well with some people more than others based on their type of depression. In other words, we have learned about the generality of the effect (the intervention) based on characteristics of the participants. Thus, between-group designs can evaluate generality directly across some condition (moderator).

Summary and Conclusions: Single-Case Experimental Research Designs

Single-case experimental designs refer to a range of arrangements that permit one to draw valid conclusions and to evaluate causal relations in a rigorous way. The designs allow experimentation with the individual subject or client. The methodology is different from the usual group research and relies on ongoing assessment over time, assessment of baseline (pre-intervention functioning), and the use of multiple phases in which performance is evaluated and altered. Three major design strategies (ABAB, multiple-baseline, and changing-criterion designs) were highlighted. The designs vary in the way in which intervention effects are demonstrated, and the requirements for experimental evaluation. Also, the designs vary in their suitability in light of practical or clinical considerations.

Data evaluation of the results of single-case designs usually relies on nonstatistical methods referred to as visual inspection. Four criteria are used to invoke this method and rely on examination of the graphed data across phases of the design. The criteria to judge the reliability of the results, i.e., whether something happened that can be attributed to the intervention include: changes in mean, level, and trend and in the latency of changes across phases. In the easy case, intervention effects are strong and these criteria are readily invoked. The data points across intervention and nonintervention phases may even be nonoverlapping, which indicates very strong effects. Yet, not all effects are so strong and invoking the criteria becomes less reliable among raters using visual inspection as the effects are less dramatic.

Statistical tests are available for single-case research. An immediate hope is that such tests would be a viable alternative to detect reliable effects when the data patterns are not obvious and visual inspection is difficult to invoke. Yet, several characteristics of single-case data, as discussed in this chapter, make it so the usual and familiar statistical techniques cannot be easily applied. Many statistical tests for the single case are available, but they are not well developed and different tests sometimes yield different results. There are exceptions, and time-series analysis was one of these discussed and briefly illustrated.

This chapter has focused on single-case experimental designs. This is the most rigorous form of the design and requires meeting and conducting several requirements, as specified by each design. The designs are useful in their own right. Yet, there often are situations where studying of the individual (e.g., psychotherapy, education) is very important but where these designs cannot be used in their most rigorous form.

Critical Thinking Questions

1. How are making and testing predictions carried out or used in single-case designs to draw causal relationships?
2. How are single-case designs different from more conventional between-group designs in assessment and data-evaluation strategies?
3. Explain how external validity or generalizing from subjects in the study to other subjects is a concern in single-case designs and in between-group designs.

Chapter 8 Quiz: Single-Case Experimental Research Designs

Chapter 9
Qualitative Research Methods

Learning Objectives

9.1 Examine the three broad influences that lead to qualitative research rising in prominence

9.2 Identify some of the data sources used in qualitative analysis

9.3 Contrast data usage in qualitative versus quantitative research

9.4 Express how the validity and quality of data in qualitative research have a lesser bearing than what they do in quantitative research

9.5 Illustrate three examples of qualitative studies

9.6 Analyze the mixed-methods research that supplements qualitative and quantitative research results

9.7 Analyze the pros and cons of qualitative research

When we discuss or consider empirical research, there is a specific methodological paradigm we have in mind. That paradigm or approach is within the positivist tradition and includes the whole package of concepts and practices (e.g., theory, hypothesis testing, operational definitions, careful control of the subject matter, isolation of the variables of interest, quantification of constructs, and statistical analyses).[1] Such is the main approach of this text, an approach where:

- One tries to devise investigations to rule out threats to validity
- Test specific hypotheses
- Identify the impact of experimental manipulations or observed variables on some outcome of interest
- Analyze the data statistically

Even the single-case designs fall into this tradition because of:

- The nature of assessment
- Specification and careful control of key variables of interest
- Methods of data evaluation

For present purposes, it is useful to refer to the dominant research paradigm in the field with the above characteristics as *quantitative research*. When people speak of scientific or empirical research, they usually are referring to quantitative research. Within the quantitative research category, there are distinctions to be made (e.g., null hypothesis statistical testing, but other less frequently used approaches as well) and we shall elaborate those later. For now, the broader category is all that is needed to convey the context for this chapter. For most researchers in the social, natural, and biological sciences, the term "quantitative research" is not used very much, because this is viewed as *the* approach or the *only* approach. Indeed, with only rare exceptions research in the premier and not-so-premier journals (or two notches below these journals where my articles often end up) are based almost exclusively on studies in the quantitative tradition.

There is another approach to research that is referred to as *qualitative research.* This is an empirical approach to the subject matter and scientific research in all of the critical ways. As I elaborate below, the tradition of qualitative research is by no means new and has deep and long roots in philosophy and the development of science more generally. Yet, the 1990s is a useful point of jumping in to convey context because of the "explosion of published work on qualitative research" (Denzin & Lincoln, 2011, p. 2). With that came handbooks, journals, and professional societies devoted to the method, as I note later. I mention these resources to convey that while we have been learning and

teaching mostly quantitative research methods (and pretty exclusively group designs), there are parallel universes both in quantitative research (e.g., single-case designs) and the broader realm of qualitative research.

9.1: Key Characteristics

9.1 Examine the three broad influences that lead to qualitative research rising in prominence

Qualitative research has its own methodology, including strategies for assessment, design, and data evaluation. Actually, in part because qualitative research encompasses so many different disciplines (e.g., psychology, sociology, anthropology, communications, education, ethnography, nursing, medicine and health care, government and policy studies), there are multiple methods and approaches.

Understandably, these approaches are detailed in various handbooks and research methods textbooks (please see For Further Reading). Although qualitative research methods are not usually taught in psychology programs, it is important be aware of these methods, what they yield, and their contribution to science. Also, the methods provide a range of options for research and researchers and yield both descriptive and explanatory information that differs from the yield of quantitative research. This chapter:

- Provides an overview of qualitative research methods in the context of psychological research
- Conveys key commonalities and similarities of qualitative and quantitative research
- Illustrates the approach through detailed examples

In addition, we take up the topic of mixed models, a formal designation that refers to the combination of quantitative and qualitative research in a single study.

9.1.1: Overview

Qualitative research is not at all new to psychology (see Banister et al., 2011; Cooper, 2012; Frost, 2011). Yet, it did not emerge from psychology. Many influences and traditions within philosophy and various scientific disciplines underlie qualitative research (e.g., Pistrang & Barker, 2012). Three broad influences can place into context the emergence of qualitative research. First, there is a tradition within philosophy that focuses on description, meaning, intentions, purpose, and context (see Packer, 2011). Approaches within phenomenology in particular (as, for example, reflected in the works of Husserl and Merleau-Ponti) provide an important starting point in light of the emphasis on description of the human experience, the role of the perceiver in understanding (constructing) the world, and such constructs as intentionality, purpose, and meaning.

The development of science and practices that we refer to as part of the scientific method in the quantitative tradition (e.g., experimental controls, operational definitions) has been marked by explicit efforts to shy away from facets of subjectivity and related internal processes and states (e.g., how one perceives, thinks, experiences, and constructs the world). Long ago mainstream psychology relied on introspection (inner reflections and reporting on one's mental processes). Indeed Wilhelm Wundt (1832–1920), the so-called father of psychology (but there were no DNA tests then to confirm), relied on introspective reports of psychological processes (memory, perception). This period of psychology included structuralism that was designed to elaborate the components (structures) of conscious experience. Individuals would systematically narrate what they were experiencing while engaging in some experimental task. The goal was to identify the basic conscious elements (e.g., descriptions such as hot, sour) that would be applicable in various combinations to all experience. Presumably, basic elements could be combined to explain and understand all of experience. Fast forwarding through various other historical periods, one can see strong efforts within psychology to focus on more objective measures (e.g., behaviors) as reflected in the works of Russian and then later American (US) researchers (Ivan Pavlov [1849–1936], Vladimir Bechterev [1857–1927], John B. Watson [1878–1958]) and a movement referred to as behaviorism. Here "objective" and observable were the main focus, and internal reporting and their foci (thoughts, feelings, emotions, sensations) were of less interest. Both structuralism and behaviorism might be considered within the positivist tradition, but the point to note here is the sharp and strong move away from reporting of the individual in internal processes as a primary source of data.

Currently many internal processes and states are studied (e.g., cognitive processes, emotion regulation), but an effort is made to move these away from the individual reports of experience. For example, within the quantitative tradition, efforts to operationalize facets of experience (e.g., stress, fear, loneliness, love, altruism) are reflected in various inventories, scales, and questionnaires.

Much has been gained by being able to measure and quantify experience in the quantitative tradition. At the same time, fundamental facets of experience are not captured by inventories and questionnaires.

This is not necessarily a function of poor inventories and questionnaires, but rather the very nature of assessment within the quantitative tradition. The level of analysis (items on a scale) is not intended to capture the richness of the experience. Rather, they are intended to provide gradations of characteristics (traits, moods, states) that can be manipulated quantitatively (e.g., measures of central tendency, variance, data analyses).

Second, within the social sciences, particularly sociology and anthropology, there has been a tradition of research in which the investigator participates in and elaborates the subject matter in great detail as a way of bringing to light key facets. Familiar examples within these disciplines can be seen from many descriptions of indigenous (once called "primitive") societies, gangs, and life in the slums (Carr et al., 2011; Lavallee, 2009). This work is qualitative in the sense that it encompasses in-depth knowledge of the people in context; participation with these cultures; rich description and narration of activities; and interpretation to place the belief, culture, and practices in context. This work is usually regarded as informative, even by quantitative researchers, and perhaps useful for generating ideas, but not usually considered as science.

Third, there has been a dissatisfaction with and reaction to quantitative research, as currently conceived and practiced. The focus on groups of individuals (e.g., mean differences between groups), the view of participants as objects of study, the investigator as an objective observer, simplification of the research situation to isolate variables, preconceived notions of what key constructs and measures ought to be, and reducing experience to quantitative results are central points of concern. In quantitative research, we are concerned primarily with how many subjects we have (sample size and power), what their responses are (means, standard deviations), and whether subjects exposed to an experimental manipulation provide different responses from those in another condition (control subjects). There are dissatisfactions with all of this even within the quantitative tradition. Yet, to qualitative researchers it is not these practices per se that are of a concern but the entire paradigm they reflect, namely, neglect of experience and attention to how individuals interpret and construct their world.

> To the qualitative researcher, the goal of understanding requires elaborating rather than simplifying the phenomena of interest; rather than control contexts (e.g., in the laboratory) and key variables, the goal is to investigate phenomena in context and as experienced by the individual. The emphasis is on the participants and how they perceive and experience the world.

The research really is participant focused—not to test some hypothesis but to learn in depth what is experienced. The experimenter is not a distant observer of subjects, but jumps in to learn close hand what is going on and must learn and get to know the subjects in a way that is more intimate than one usually considers in research. To understand behavior, key concepts such as meaning and purpose, usually avoided in quantitative studies of human functioning, are central topics within qualitative research. The broad variables including context in which "variables" operate, perceptions, goals, and interactions are the central foci of qualitative research.

9.1.2: An Orienting Example

We will certainly clarify the approach more formally and contrast qualitative and quantitative research, but consider a brief example that will orient us to what qualitative research has as its foci and how those contextual points I have highlighted relate to real people.

At a clinic where I work, children are referred for treatment because of their extreme aggressive and antisocial behavior (Kazdin, 2010). We begin with an extensive set of assessments that address multiple facets of the child, parent, family, school, and more. Among the domains we measure is parental stress because it is related to several facets of treatment participation, often is a significant issue in its own right as an antecedent, consequent, or correlate of child functioning, and is an outcome measure that changes when treating the child. We use a few measures of stress; they have subscales to assess types and sources of stress (e.g., related to sources generated by the child and those related to the parent's own life). The measures include all the wonderful psychometric properties related to reliability and validity, have been well studied (not by us), and are useful. That said, the measures leave a lot out that qualitative research would capture.

At our clinic, occasionally we have mothers who are waiting for their husbands or significant others to return home after a prison term. While the partner has been in prison, sometimes for relatively brief periods (e.g., 3–12 months), the mother may have begun a new relationship with a live-in significant other and/or has become pregnant with someone else's child, or occasionally has already delivered that child. The mother may live in great fear of her life as the date for her husband's release from prison approaches. Now that we have seen this on a few of occasions, it is really difficult to imagine any psychological measure of stress to even approximate the stress, dread, and terror these mothers' experience, as they have voiced to us. Yes, of course, this stressor could be operationalized and measured and one could imagine developing a scale with a bunch of questions (e.g., each on a 5-point scale, where 1 = not at all and 5 = greatly), such as "To what extent do you fear your life is threatened by the release of your husband from jail?" Probably no researcher in the quantitative tradition would think for a moment that a high score (5) captures the intensity, duration, or scope of stress and angst. Almost certainly, no mother would. This level of quantification misses so much of the stress the mothers' experience in the circumstances I have highlighted. Perhaps we could add to self-report a set of biological measures that show the many ways stress is evident physiologically (e.g., hormone release, brain activation). Piling on more wonderfully objective measures (in the quantitative tradition) is still missing much of the point.

Most of us have not experienced this type of partner relationship and stress. Something closer such as the death of a parent, sibling, or close relative is a more common illustration of the depth and scope of the experience. What measure of pain, loss, and emotion could capture the experience, and how are we changed by it? And while many of us are changed in ways that might be similar, we are also changed in ways that might be different. To be sure, there are objective measures and quantitative indices. Yet, qualitative research is designed specifically to reflect the richness, depth, and meaning, and complex situations in which they emerge and would be an approach to better capture the experience of stress of the mothers I mentioned and of the loss of a significant other that death reflects. We do not only want to see how super-stressed individuals express their stress and dread on our objective measures based on fixed and predetermined views of the experience (questionnaire items), but we want to hear the voices of the individuals themselves and their experience of stress. And of course we want to do this scientifically and not just collect anecdotal reports. Qualitative research is at the level of analysis to get at experience in the ways I am indicating. That requires directly participating with and engaging individuals to obtain the details and full range of the experience. We can go beyond standard questionnaires and capture the intensity of the experience of mothers I have described and obtain the range of emotions, the similarities, and the individual differences. (Pretty intense chapter so far!) Let us now begin more formally to discuss the approach.

9.1.3: Definition and Core Features

Qualitative research is an approach to the subject matter of human experience and focuses on:

- Narrative accounts
- Description
- Interpretation
- Context
- Meaning

The goal is to describe, interpret, and understand the phenomena of interest. Through description and interpretation, our understanding of the phenomena can be deepened. The process of achieving this goal is to study in depth the experience of the participants, i.e., those who are studied, and to convey how that experience is felt, perceived, and the meaning it has for those whose experience is being presented.

As a general rule, qualitative research relies heavily on description and interpretation of the experience or action that is studied. The purpose is to describe the experience (e.g., thoughts, feelings, and actions of a person in a particular context; interactions in which people engage) but to do so in a way that captures the richness of the experience and the meaning it has for the participants.

> We say that qualitative analysis increases our understanding because the level of analysis is detailed and in-depth and ideally brings to light new ways of speaking about the phenomena that the investigator, reader, and indeed participants may have not fully understood prior to the analysis.

As a way of describing this more colloquially, we know about many of life's experiences (e.g., being worried, being infatuated, being in love, experiencing stress). A qualitative analysis is designed to bring these experiences to light and to make the tacit explicit in ways that provide a deeper and more empathic understanding.

All of the above may appear too fuzzy and so permit a brief time out and quasi-digression. Any reader who has been trained in the scientific method (i.e., the quantitative tradition) ought to experience a little discomfort (or a panic attack) at this point in the discussion. Perhaps qualitative research sounds loose, laced with subjectivity, and riddled with precisely those problems (e.g., subjective interpretation of the investigator) that current scientific methods were designed to redress. Moreover, qualitative research may sound like writing or reading a good textbook or hearing a good case study, each of which may be a richly detailed account of some facet of human experience.

> *Qualitative research very much relies on the traditions captured by literature, namely, descriptive accounts that elaborate experience.* Yet, qualitative research provides a systematic, empirical, and scientific approach to description and understanding and provides replicable, reliable, and valid accounts, although these terms have somewhat different meanings.

That is, systematic and scientific tenets and procedures of qualitative research move it out of the realm of literature or other arts where experience is also portrayed. Mind you, there is nothing wrong with literature and I have caught one of my friends even reading some once in a while. Yet, qualitative analysis goes beyond this by providing systematic (but not usually quantitative) methods of data collection, analyses, replication, and efforts to address biases that can influence the data (Denzin & Lincoln, 2011).

9.1.4: Contrasting Qualitative and Quantitative Research

Normally it would be useful to contrast two approaches after fully describing and illustrating qualitative research. Yet, qualitative research can be brought to light by juxtaposing key characteristics to those of the very familiar quantitative research. In terms of broad goals, qualitative and quantitative research both seek to understand natural

Table 9.1: Select Characteristics that Distinguish Quantitative and Qualitative Research

Characteristic	Quantitative Research	Qualitative Research
Goals	Test theory and hypotheses; identify causal relations, seek group differences or patterns	Describe and interpret experience; provide new insights, describe and explain with few or no initial hypotheses; generate theory
How to Study	Isolate variables, control potential artifacts and extraneous influences; rule out rival hypotheses	Consider variables as they appear in context with all of the natural influences; complexity is embraced to elaborate the gestalt as well as any key influences in context
Subjects	Study (or try to study) a large number of subjects for statistical power	Study one or a small number of cases (individual, culture, organization) intensively
Use of Control Conditions	Usually control or comparison groups are included to address threats to validity (e.g., internal, construct)	No control group. The goal is to elaborate the richness of a particular group and the commonalities and differences that may emerge within that group
Role of the Subject/ Participant	The subjects are the object of study, the people who provide the data; the subjects do not reflect on the data or help the experimenter make sense out of the results	The participants are not objects; the experimenter and subjects become one in the sense that the experience described and understood cannot be removed from the one who describes (subjects) and the one who extracts meaning and themes from that (experimenter); the subjects are often consulted to ask whether the description and interpretation capture the experience
Role of Investigator	Minimize the investigator's role; the perspective, views, and feelings of the investigator are reflected in the hypotheses or focus of the study, but not in the methods; the investigator is detached to the extent possible, so the findings can stand on their own without her or his involvement	The investigator is better referred to as another participant and part of the interpretation in light of his or her perspective; the perspective is made explicit, but it can never be removed; empathy of the investigator is encouraged as a key to deeper understanding; the investigator is engaged rather than detached and can understand better to the extent that meaning of the situation is experienced
The Data	Scores on measures that operationalize the constructs; standardized measures are used whenever possible; the data refer to information that has been reduced to numbers	Narrative descriptions, full text, lengthy interviews, accounts, examples; the "story" details the subject matter in the context of how it unfolds, happens, and is experienced; the "words" are the data and are not reduced to numbers
Data Evaluation	Statistical analyses to find patterns, averages, to control influences further, to identify the impact of variables on each other and on an outcome	Literary, verbal, nonreductionist, go from description to interpretation to identify themes to bring new qualities to light. Systematically identify themes and ways of categorizing experiences for purposes of presentation
Criteria for Knowledge	Procedures and findings can be replicated	Descriptions are coherent and viewed by others (colleagues, participants) as internally consistent capturing the experience; procedures and findings can be replicated within the study itself (e.g., by confirmation of another investigator) and as well as in further additional studies
A Major Contribution	A new theory, hypothesis, or relation is brought to light that will increase our understanding of the phenomenon	Our understanding of the experience is elaborated and brought to light in depth as well as in ways and that extend our understanding. Developing theory grounded in close study of the phenomena of interest is a special strength

NOTE: As a general rule, I personally object to tables that contrast approaches with two columns because the structure implies qualitative (i.e., categorical) differences, emphasizes extremes, and fails to consider the inevitable fuzziness of many of the distinctions. The value of the table is to draw sharp lines to introduce the approach, but it is not difficult to identify a study in one tradition (i.e., quantitative or qualitative) and show how many of its features are captured by the columns of the table designed to characterize the other tradition. Later in the chapter, we will discuss mixed methods that combine qualitative and quantitative approaches.

phenomena, to provide new knowledge from systematic methods, and to permit the findings to be replicated by others. Several key differences in how the subject matter is approached help to convey those special characteristics of qualitative research. Table 9.1 provides salient dimensions of how one approaches research and the differences in quantitative and qualitative approaches.

9.1.5: More Information on Contrasting Qualitative and Quantitative Research

Clearly the key difference is in the specific goals of quantitative and qualitative research because from these goals the different methods and foci naturally result. Research in the quantitative tradition tries to understand a phenomenon and focuses on explaining the underlying processes, the causes, and how the phenomenon occurs. For example, a quantitative evaluation of suicide would include understanding the risk factors, including individual, family, and contextual features and how these unfold to lead to suicide. The genetic and neurobiological underpinnings (e.g., neuroimaging, neurotransmitters) of individuals prone to suicide and in an induced or natural state of hopelessness, a key characteristic of suicidality, too would be well within the quantitative research tradition. A causal explanation is sought to explain how and why suicide occurs and who is likely to be at risk. The data for the research are likely to come from standard questionnaires (e.g., hopelessness, depression) and measures of other assessment modalities (e.g., cognitive processing computer tasks, neuroimaging, biological indices of stress). The measures

would be scored, and sums of scores from each measure would be used to characterize the group or subgroups. The data might lead to predictors of suicide, perhaps subgroups, mediators, and moderators, and so on. Indeed, there are remarkable gains in our understanding from the quantitative tradition (e.g., Hawton, Casañas i Comabella, Haw, & Saunders, 2013; Nock et al., 2013).

Now shift to the qualitative tradition. Understanding focuses on how individuals experience suicidal ideation and attempt and their thoughts, feelings, actions, and how they construct their lives in a way that makes suicide an option, understandable, or a viable way of thinking. Here a score on measures of suicidality and related constructs do not capture suicidality of the individual as experienced, a point made earlier in my discussion of mothers and stress. Indeed, many items on all of the usual measures of the quantitative tradition may not be relevant to a particular individual's condition, many items that might be relevant are not given the weight they play in any particular person's experience, and many relevant domains may not have items on the measure at all. A standard questionnaire with all its enormous strengths in the quantitative study is a constraint and possibly not relevant in a qualitative study.

> Qualitative research seeks to understand action and experience and hence must view broad sweeps of functioning (affect, cognition, behavior) in context and then obtain the detailed descriptions within each of the areas.

Apart from studying the complexity of experience in its full bloom (i.e., as it unfolds in uncontrolled everyday situations), the qualitative approach views the investigator quite differently from how he or she is viewed in the quantitative approach. The model in quantitative research is that of an objective scientist, looking through a telescope; the goal is to have the investigator serve as an objective instrument, and indeed it is even better if the material observed through the telescope can be recorded in ways that minimize the influence or active role of the investigator. Quantitative research methods in psychology and social sciences more generally have been very much modeled after the natural and biological sciences (e.g., biology, physics, and astronomy) and keeping the investigator separate from the subject matter is axiomatic.

> In qualitative research, the investigator is not someone who "collects data," but rather someone who participates with the subject to bring out the data and then integrates the information in a way that affects the data, i.e., gives it meaning and substance.
>
> Elimination of a frame of reference or perspective of the investigator may not be possible and if possible, not necessarily desirable. In fact, to really understand the phenomenon, many qualitative investigators become deeply involved in the subject matter; understanding is optimal when one experiences the phenomenon directly, intensely, and close up. Thus, studies of other cultures that have provided the greatest insights often stem from those who enter and live in the culture for an extended period. Pertinent to clinical psychology, many excellent insights and hypotheses have come from intense and close evaluation of individual cases, as noted in the discussion of the source of ideas for research.

9.2: Methods and Analyses

9.2 Identify some of the data sources used in qualitative analysis

The basic information (data) used for a qualitative study can be obtained in many different ways. Among the salient methods and sources are:

- Interviews
- Direct observations
- Statements of personal experience
- Documents (e.g., personal journals, diaries, letters, biographical materials stories passed across generations)
- Photographs
- Audio, or video recordings
- Films

Each of these methods has its own recommended approaches and options for data collection (see Denzin & Lincoln, 2011).

9.3: The Data for Qualitative Analysis

9.3 Contrast data usage in qualitative versus quantitative research

At first blush, these methods do not look all that different from many measures and ways of collecting data in quantitative research. For example, direct observation and interviewing also are used in quantitative clinical research.

> In qualitative research, direct observation is more naturalistic, i.e., it occurs in naturalistic contexts as the subjects would normally participate and interact. The investigator is drawn into the world of the subjects (i.e., the family, playground, school), and predetermined categories (to code behavior) usually are not used.

In contrast, within the quantitative approach, much of the direct observation consists of standardized assessments in which the world of the subjects is brought into the lab (e.g., observe marital interaction on standardized tasks) or the world of the investigator is brought to natural settings

(e.g., observations conducted in home with efforts to standardize the situations that govern family interactions). In both instances, the investigator knows the codes to operationalize constructs of interest. Starting out with assessment codes and the constructs of interest are essential to quantitative research. Typically, *not* starting out with assessment codes and the constructs of interest is essential to qualitative research.

Information obtained from the sources (e.g., interviews, observations) initially is taken as descriptive material and serves as the basis for analysis. Analysis of the information takes many different forms to:

- Look for recurring themes or key concepts that emerge in peoples' descriptions of their experiences
- Identify processes or a progression that seems to show the flow of experience
- Link variables that emerge concurrently or over time
- In general look for consistencies and patterns in the material[2]

One of the advantages of the approach is to discover and to generate new conceptualizations of phenomena in the process of bringing special features to light. Although methods of evaluating the material to pluck themes can be time-consuming, computer software is available that is designed to facilitate making these connections and interpretation from qualitative data by displaying the information and pointing to possible patterns.[3]

Occasionally, the approach to cull meaning and to examine connections among variables utilizes methods of quantitative analyses. For example, in some qualitative studies, the descriptive material is coded into categories and analyzed statistically to examine the occurrence of themes and their interrelations, although these are exceptions. As we see later, sometimes the methods of qualitative and quantitative analyses are combined in mixed-methods studies.

9.4: Validity and Quality of the Data

9.4 Express how the validity and quality of data in qualitative research have a lesser bearing than what they do in quantitative research

In quantitative research, different types of validity were distinguished (internal, external, construct, and data evaluation). These are all related to drawing inferences about experimental manipulations or observed conditions and in particular on drawing causal relations. Qualitative research does not have the same goals about evaluating the impact of independent variables in quite the same way. Consequently, the specific types of validity and their threats do not transfer easily or directly to qualitative research. For example, selection bias, statistical regression, diffusion of treatment, novelty effects, weak statistical power, and so on are not very relevant.

9.4.1: Validity

Within qualitative research, the concerns about obtaining findings that are reliable (i.e., consistent and replicable) and valid (i.e., reflect the phenomena of interest) are no less than they are within quantitative research. Yet, there is no universal list of types of validity (Onwuegbuzie & Leech, 2007; Whittemore, Chase, & Mandle, 2001). In Table 9.2, I summarize five types of validity that qualitative research tries to address and that are widely recognized among researchers who conduct such research. I use the term "validity" to help draw connections between qualitative and quantitative research. The terms in the table convey the thrust of qualitative research in yet another way, namely, what it is trying to accomplish and how it will ensure that the data can be trusted. Yet, "validity" is not commonly used in qualitative research as I mention next.

Table 9.2: Unique Types of Validity that Convey Features of Qualitative Research

Types of Validity	Features
Descriptive Validity	The extent to which the account reported by the investigator is factually accurate. The account may reflect descriptions of events, objects, behaviors, people, settings, times, and places.
Interpretive Validity	The extent to which the meaning to what has been described is accurately represented. Is the descriptive material understood adequately? Are the views, intentions, feelings, or other data given an account interpreted in a way that represents or understands the experiences?
Theoretical Validity	If explanations are designed to address how and why a phenomenon or experience has occurred, how well does the explanation "fit" the data? The theory is more abstract and at a higher level of inference than interpretive validity and conveys the possible reasons or underpinnings of the phenomenon.
Internal Validity	Similar to the notion in quantitative research, are there other sources of influence that could explain the results apart from the influence the investigator identifies?
External Validity	Also, similar to the notion in quantitative research, are the findings generalizable across people, time situations, and settings?

9.4.2: Qualitative Research on and with Its Own Terms

Qualitative research has its own terms used to capture critical features of scientific inquiry, namely, that the methods are systematic and consistent in what they yield and

capture the phenomena of interest. These serve the purposes of reliability and validity. For easy reference, key terms and concepts are included in Table 9.3. A key concept on which validity depends is triangulation.

Table 9.3: Key Concepts of Qualitative Research to Ensure Consistency and Validity of the Data

Concepts of Qualitative Research	Description
Triangulation	The use of multiple procedures, sources, or perspectives to converge to support the conclusions. Triangulation may rely on separate bits of data, different methods of qualitative analyses or qualitative and quantitative analyses of the same data, use of different theoretical frameworks, and different investigators.
Confirmability	The extent to which an independent reviewer could conduct a formal audit and re-evaluation of the procedures and generate the same findings. The extent to which results are confirmable by others depends on the care with which the original investigator conducts the study to begin with and the methods of triangulation used for the demonstration.
Credibility	The extent to which the results are believable. Would the descriptions provided by the investigator be viewed as credible by the participants and by others who have also had that experience but were not included in the study?
Transferability	The extent to which the findings are likely to be generalizable or limited to a particular context (are context bound). This is evaluated by looking at any special characteristics (unrepresentativeness) of the sample and identifying the likelihood or plausibility of extending the findings to similar circumstances.

Triangulation refers to using multiple procedures, sources, or perspectives to converge to support the conclusions.

Triangulation may rely on separate bits of data, different methods of qualitative analyses or qualitative and quantitative analyses of the same data, use of different theoretical frameworks, and different investigators.

When different ways of examining the problem or phenomenon converge in the information they yield, this is triangulation and strengthens (better establishes) the validity of the finding.

9.4.3: More Information on Key Concepts and Terms

Triangulation can be achieved in different ways:

- The investigator can use multiple sources of data (e.g., interviews and questionnaires to combine qualitative and quantitative methods).
- The conclusions can be examined by others both to examine the descriptions and interpretations.
- Often the participants themselves are asked to reflect on the data and interpretations of the investigator to see if they concur or have information to add or alter.
- Peers (other colleagues) might be involved to challenge the interpretations and conclusions. Often investigators of the study will independently develop or evaluate the interpretations as a check, but peers not involved in the study may participate in this role as well.
- Finally, considering cases that seem to disconfirm the researcher's conclusions is another check.

Counter instances may not refute the explanation but may provide other explanations that ought to be considered, that might apply to others, or raise themes about the experience that were not previously considered. Indeed, when capturing the experiences of individuals, there is no "counterinstance" per se—one is not looking for a homogeneous way to characterize everyone, and varied experience is central to the approach and subject matter.

Triangulation is used to bolster the strength and validity conclusions and addresses many of the issues noted in Table 9.2. Essentially, in more familiar terms, triangulation resembles multimethod assessment approaches in quantitative analyses. In quantitative research, the strength of conclusions is usually bolstered by using more than one measure of a construct (e.g., depression) and more than one way of assessing that construct (e.g., self-report, other report, implicit attitude test of affect words). In qualitative research, triangulation covers more than multiple assessment methods, and can include multiplicity or generality across many different aspects of the study, including the range of participants (investigators, subjects) used to derive conclusions. In general, as with quantitative research, there are methods within qualitative research to address:

- Alternative interpretations
- Bias and artifact
- Replicability of the results

Confirmability refers to the extent to which an independent reviewer could conduct a formal audit and re-evaluation of the procedures and generate the same findings.

The extent to which results are confirmable by others depends on the care with which the original investigator conducts the study to begin with and the methods of triangulation used for the demonstration.

Confirmability of course reflects replicability of findings and is central to all scientific research. Yet, replication in quantitative research usually refers to a separate study that is completed to test whether results of an original study can be repeated. Replication in that sense also applies to qualitative research. Yet, confirmability is a practice within a given qualitative study to validate the description and interpretations that were made by someone else who evaluated the information (e.g., interviews, diaries).

Credibility or believability of the results is another check on validity. *Credibility addresses the question of whether the descriptions provided by the investigator would be viewed as believable by the participants and by others who have also had that experience but were not included in the study.*

In everyday life, credibility in this way is occasionally experienced when a friend mentions something and we recognize this immediately as exactly the feeling or experience we had. In qualitative research, coherence of the interpretation, agreement among others about that interpretation (including when possible the participants themselves), and consensus that our understanding of the experience or phenomenon is enhanced as a result of the analysis are salient among the criteria to evaluate the findings. Does the analysis capture the experience and extend our understanding (e.g., of the experience of living with HIV/AIDS, of growing older, of being a child, of living in a particular culture, of serving as a prisoner of war)?

Transferability pertains to whether the data are limited to a particular context (are context bound) and is evaluated by looking at any special characteristics (unrepresentativeness) of the sample.

Transferability is related to generality of findings and hence external validity, as discussed in quantitative research, but warrants separate consideration.

To permit evaluation of transferability, investigators provide detail of the context in which the study was completed (e.g., special features of the participants, setting, experience, or event) to allow the reader of the report to decide whether the circumstances resemble and are likely to apply to another context that is similar.

As in quantitative research, generality has to be tested directly as well as making inferences about the likely applicability across situations.

9.4.4: Checks and Balances

Understandably, in qualitative research, there is the concern that the views of the investigator may play a particularly significant and unchecked role in the interpretation. That is because the investigator often participates directly with the participants, collects extensive information (e.g., hours of interviews), and now pours through that to cull themes and consistencies. The concern is that perhaps the consistencies are in how the investigator views the material or normal cognitive biases that any human would bring to the situation might unwittingly place preconceived and even culturally bound structure to the information. Although the perspective of the investigator is important, as consumers of research we do not accumulate findings that are only pertinent to or generalizable to the particular investigator.

Multiple strategies are used to help ensure that the data are not mere reflections of the investigator's perspective.

The strategies used are:

1. Investigators are encouraged to make explicit their own views, including how their expectations may have been met or not, what was consistent and discrepant from any preconceived views, and what orientation or approach they may be taking to the subject matter (e.g., observing it from a particular perspective or even theoretical orientation, discipline or frame of reference). Noting this perspective, orientation, and expectation permits others in the scientific community to evaluate the interpretations in light of potentially important influences.
2. There is an iterative process (i.e., repetitive, checking) in which investigators are encouraged to consult with other investigators to identify the extent to which the raw materials (e.g., lengthy narratives, audio or video taped materials) are likely to reflect key themes the investigator has identified.

Are the interpretations cohesive and do they capture the experience?

What do you think?

These are questions posed to others who evaluate the qualitative material. Sometimes the participants also are part of this verification process. During the process of collecting the information or after the information is collected, participants are encouraged to review the categories, broader concepts, sequence of experiences that have been proposed and to make comments. These comments themselves are brought into the process to elaborate, refine, or alter what has been proposed by the investigator. Thus, the fact that the investigators are people who have their own perspectives, experiences, and shaded glasses does not doom in any way the resulting data to an idiosyncratic perspective. There is a consensual process involving other investigators as well as participants.

Other investigators evaluate the process of reaching the interpretation by examining the procedures, raw data, and analytic strategies as well as the conclusions themselves (see Miles, Huberman, & Saldaña, 2014). The investigator is encouraged to make raw data available to others during the investigation to permit a check on how the information (e.g., transcripts) reflects themes that have been identified. Thus, there is an internal replicability that is part of the evaluation, i.e., scrutiny by others. All of this is facilitated by procedures and computer software alluded to earlier that help to display, code, systematize, and retrieve the data and to test the emergence of broader constructs and categories that form the basis of the investigator's interpretation.

As evident in this discussion, concepts and principles underlying qualitative research are similar to those of quantitative research, namely, are the findings reliable and valid?

Reliability pertains both to:

- The methods of studying the data (e.g., how themes and categories are identified, how interpretations are made)
- Coherence or internal consistency of the interpretations

Validity refers to the extent to which there is a finding that makes sense, captures experience, and is confirmed and is independently confirmable by others. And of course, keys to scientific knowledge obtained by quantitative or qualitative approaches are in replicability (can the procedures be replicated) and then actual replication (are the results obtained by others evaluating similar circumstances).

9.5: Illustrations

9.5 Illustrate three examples of qualitative studies

Although qualitative studies are available in psychological research journals, they are not that common.[4] Add to that the fact that the topic is not usually taught in undergraduate or graduate work in clinical psychology. As a result, it is worth conveying a few examples in detail to convey the foci, how one goes about the research, evaluates the data, and reaches conclusions. Consider three examples on very diverse topics, including:

- Surviving a bus crash
- Being subjected to discrimination and verbal and physical harassment based on one's sexual identity
- Regretting something one has done on Facebook

9.5.1: Surviving a Major Bus Crash

Overview: Bus crashes occur frequently and can result in serious injury and death. The experience of surviving a crash or other such disasters like that obviously can be life-changing. We know all too well that exposure to natural (e.g., tsunamis, hurricanes, tornados) or person-made catastrophes (e.g., war, school shootings, terrorist acts) can lead to posttraumatic stress disorder (PTSD), just to mention one of the consequences.

In this study the authors examined survivor experiences following a bus crash in Sweden (Doohan & Saveman, 2013). Two busses crashed almost head on a snowy narrow road; six passengers died. The 56 survivors (18–64 years of age) included individuals with varying degrees of injuries. Approximately one month after the accident, extensive detailed interviews were obtained from the survivors. Interviews were recorded and transcribed (about 1–3 pages per participant). The goal was to characterize the experiences in a systematic way to examine the domains of relevance to the participants and the range of their reactions.

Qualitative analyses of the full text focused on content analysis, i.e., themes that emerged by coding statements. The narratives were reviewed (by two individuals), and categories and subcategories were identified.

Supportive and illustrative quotes were used to clarify the categories. Themes that emerged as the categories/subcategories were:

- Feeling, thinking, and helping others
- Reaction to the emergency care
- Encountering the media
- Receiving formal support
- Healing with social support (family and friends)
- Difficulties in sleeping
- Everyday traveling (problems)
- Seeking closure

Within these themes, the specific reactions of the survivors were further described, only a few of which are sampled here from various categories. The results indicated that participants felt shocked and lost, and were delirious after the crash as if they were spectators rather than victims. Their initial thoughts were to call their families and friends. Helping acts and support by other victims were identified as important. For those victims who did not help others, even if their physical condition prevented that, feelings of guilt and anguish were expressed.

Passengers felt negative experiences associated with media at the crash site and after the crash at their homes or neighborhoods. They felt the media were intrusive, scary, and unprofessional. The specific unpleasant and distasteful experiences included:

- Being filmed and photographed
- Being interviewed while they were in shock
- Having the interviews published without their permission

Support from emergency workers was viewed as helpful. In the days after, professional help (e.g., mental health professionals, support groups) was satisfactory to many but there was great variability. Some did not want any support from professionals at first but did later. A few days after the crash, the supports from family members, friends, neighbors, and fellow passengers were important sources of support. Even with strong support, many participants experienced symptoms of PTSD (e.g., flashbacks) and had sleep difficulties. Most of the participants were uncomfortable with traveling in any vehicles or when walking near traffic. Passengers expressed interest in knowing more about the cause of the crash; some went to view the crash

site and with the support of other passengers felt there was a benefit to help bring closure.

> The very nature of qualitative research means that summaries such as the type I have attempted cannot represent the richness of the data.

Means, standard deviations, and various statistics usually used to summarize the results in quantitative studies are not the currency of qualitative research precisely because one does not usually want to blend, combine, aggregate data too much, and in the process lose the varied individual experience. In presenting the results, the categories identified were enriched in the original report by direct quotes to illustrate the themes and varied reactions.

9.5.2: Comments on This Illustration

> **Comment:** So what have we learned from this, and how is this qualitative study helpful or of use? (In fairness, this same question is reasonable to ask of any quantitative study.) Among the contributions, the categories or themes that were identified convey consistent domains that capture pertinent experiences and reactions. The categories themselves provide insights about what facets of experience are likely to emerge immediately and over time and what might be done differently or better to address negative reactions within the categories.

Within the categories, there was some variability in reactions, but some of the commonalities were important to identify. The diverse types of support all seemed to be important immediately after the accident (by other victims who could help), by emergency workers, perhaps less so by professional health workers, and then in the ensuing days by family, friends, and neighbors. Family members seemed to be the main source of support and viewed as most helpful. This finding might suggest one place to intervene better to minimize the adverse reactions—bring in family and support as soon as feasible. Also, instructing and helping family members in ways they can provide support also might decrease stress and improve or speed up recovery of the victims. The findings suggest the benefits of protecting victims from the media that exacerbate negative reactions to the crash. In short and again only highlight some of the findings, we know likely areas to address to help and protect victims.

> The qualitative study provides information that would be unlikely to emerge from a quantitative study.

Quantitative research is more likely to use questionnaires with predetermined items and scales. Those scales (e.g., depression, trauma symptoms) would be reliable and valid measures, so there is no criticism on that score. Yet, to identify emergent themes we want to hear all that survivors wish to say and to use reliable methods to identify themes and specific reactions. There is no need to pit qualitative and quantitative research against each other. In fact, the results of this qualitative study would be a good basis to develop new measures that reflect reactions and recovery from accidents. These measures could be used in quantitative research to describe, predict, and evaluate who is likely to have what kind of reaction to accidents.

Although I have only presented highlights, it is clear that we have learned some interesting things from this study that not only elaborate the experience of the crisis but also suggest how we can be more helpful in getting people through such events. And the results might be transferable (generalizable) across many disasters that people experience. Among the areas for further qualitative research would be to evaluate common themes across different type of disasters.

9.5.3: Lesbian, Gay, Bisexual, and Transgender (LGBT) Youth and the Experience of Violence

> **Overview**: LGBT youth usually become aware of their attractions and gender identity between ages 12 and 14, although the age is slightly older for transgender identity (15–17). There may be many other reasons to distinguish among LGBT, but the different subgroups lamentably share some untoward experiences.

These individuals experience high rates of harassment directly (e.g., bullying) and indirectly by hearing the use of insensitive terms (e.g., gay, fagot) not directed at them but that make the environment hostile and unfriendly (Kosciw, Greytak, Bartkiewicz, Boesen, & Palmer, 2012). Words are harmful enough, but LGBT youth often experience high rates of physical harassment and assault as a result of their sexual orientation or identity. Victimization rates can be very high (e.g., up to 80% of individuals, as reviewed in the study I highlight here [Grossman et al., 2009]). The consequences are enormous. The physical and verbal harassment makes the youth feel unsafe at school and their school performance deteriorates (e.g., as reflected in missing classes and entire school days). The harassed victims are more likely to get into physical fights, to attempt suicide, and to use illicit substances. We also know from studies on bullying that the risk of serious short-term and long-term effects on mental health is enormous (e.g., Copeland, Wolke, Angold, & Costello, 2013; Turner, Exum, Brame, & Holt, 2013). These and other such findings come from quantitative studies that document prevalence, risk, consequences, and other critical features of LGBT exposure to harassment and violence. But what do the victims have to say? What are their experiences, perceptions, thoughts, and feelings through all of this? The goal of the study was to understand what LGBT youth found as oppressive and destructive conditions in the school.

This study comprises 31 LGBT youth (aged 15–19) from minority groups, including (and in order of their proportions in the study) "mixed race," African American, Hispanic/White, Hispanic Black, and Asian. Youth volunteered to participate and were drawn from 21 different New York City public schools, mostly of ethnic minority in an urban school environment, and met in focus groups to talk about their past and present experiences of school violence. The focus group format was selected to foster exchanges among participants and to help them clarify the meaning and behaviors of the experience, with the idea that hearing and talking would bring out the material more richly. The groups were small (5–8) and varied in composition (e.g., all five male to female transgender youth in one group). Each group was led by a facilitator who had prior experience in working with LGBT youth. The activity lasted up to 2 hours. Several questions guided the groups, and both positive and negative experiences were discussed.

The group conversations were audiotaped and transcribed. The transcribed material was analyzed by looking for themes using a method (techniques of grounded theory) that guides identification of material by specific methods beyond the present scope. The material was independently coded and corroborated by two evaluators for trustworthiness, a concept mentioned previously.

Three broad themes emerged:

1. Core themes of the groups
2. Being subjected to negative attention
3. Recommendations for prevention

Within each, details and quotes were used to elaborate subthemes that were evident. For example, under core themes, LGBT youth felt:

- They had little control over school violence.
- They had no sense of being part of the school community.
- They felt that not much could be done to remedy the situation.
- They felt that heterosexual youth had a perceived and actual sense of power over sexual minority youth.
- Rarely did others (e.g., teachers, administrators, guards) intervene to help them.

In the negative attention theme:

- Youth reported being objects of hate speeches, name calling, insults, and harassment and felt they had to always be on guard.
- Some delineations of subgroups were noted; youth believed that lesbians had an easier time than gay males who were self-identified.
- In the theme, recommendations for prevention of verbal and physical attacks, youth underscored the importance of being truthful about themselves and not trying to pass for something they were not.
- They identified the responsibility of the harassment on individuals (perpetrators) and did not believe the schools had responsibility for the problem.

Yet schools could help by making the overall environment much more sensitive to LGBT issues. LGBT speakers and organizations could be brought in or formed, and school personnel could be alerted to the impact of even hearing indirect pejorative references or comments that constitute harassment. Finally, school personnel who were LGBT and who "came out" would be very helpful to serve as role models at school and as part of a larger educational effort to convey that LGBT status is not something wrong or to be hidden.

9.5.4: Comments on This Illustration

Comment: Is there any value to what we have learned? First some context. The youth were multiracial and of ethnic minorities. These latter groups alone are subject to much higher rates of harassment and bullying leaving aside their identification with LGBT. Related, the group minority-LGBT youth is not likely to be studied very extensively in qualitative or quantitative research. Accumulating a sufficient number of participants alone is not easy. Overall the study included a relatively neglected group that is likely to experience high rates of harassment and negative experiences. This study conveyed the feelings and plight of LGBT students.

The students related in detail how they were victimized but also how they felt powerless in exerting control. They also felt unsafe at school and always had to be on guard, understandably in light of the direct and indirect harassment. The lack of community (belonging to the group) and agency (power to control their environment) was the most salient theme that emerged.

This overview cannot provide all of the details from the individual stories and quotes that serve as the primary data. The individual details are not trivial and indeed could be the basis for strong action. For example, one student noted that in his school with ethnic minorities, the "n" word could not be used (in relation to African Americans) and the "f" word could not be used in swearing. (I am making much of this here based on my own reaction rather than points the authors promoted.) I have no reason to suggest that the intervention—prohibiting use of some words—alone makes a difference, but the schools' silence on LGBT slurs comes to the fore in seeing what does and does not "count" as improper and completely inappropriate language. Silence here could easily be seen as tacit if not direct support for harassment of LGBT youth or minimally as not elevating the issue to the plane of other related issues. Either way

this is one of those cliché moments where the "silence speaks volumes." If nothing else, the study should prompt self-scrutiny and change within school administration and policy.

More generally, the qualitative study adds details that we would not otherwise have. The information could be used in helpful ways. For example, in educating teachers and administrators, the last thing likely to help would be a list (or endless PowerPoint slides) noting statistics about who is harassed and what the effects are. That quantitative information would be a great starting point. However, perhaps the suffering and experience of individuals and the results of the focus group with real people and real quotes have a much greater likelihood of effecting change. Essentially, the qualitative results say, "in your school, children feel horrible, feel no support from you, and are suffering psychologically and physically, moreover the school could do more." That message with qualitative details might help move toward genuinely helpful interventions. Some of these interventions were even suggested by LGBT youths themselves.

9.5.5: Yikes! Why Did I Post That on Facebook?

Overview: Social networking has so many benefits in keeping in close and even moment-to-moment touch over both short and long distances. And networking continues to grow in numbers and formats. As of 2010, over 600 million users of Facebook have made that the largest social networking site. There are some downsides of the ease of posting and commenting that Facebook and other such opportunities allow.

Self-disclosure and photos, for example, may provide troublesome fodder when they reach the hands of peers, colleagues, students, and employers. Indeed, there are scores of documented instances where people (e.g., U.S. Congressman, teachers, administrators) have lost their jobs or were forced to resign based on postings of photos or comments (e.g., Smith & Kanalley, 2011; Warran, 2011). There are probably many more undocumented instances in which offers for jobs, scholarships, awards, and other benefits have not been provided based on Facebook and other social media revelations. Understandably, major organizations (government agencies, schools, companies building a positive community image) have to sustain their standards and images. Having someone with a Facebook page that reveals racy images, wildly radical political views, or a disinterest in methodology is not worth the risk of being tarnished.

As individuals we realize this and once in a while send out something we wish we had not sent. In this qualitative study, the focus was on the experience of regret among Facebook users who wished they had not sent something (Wang et al., 2011). The goal was to identify what users regret that they posted, the causes of the regret, and leads for developing countermeasures so that users could better avoid problems in the future. This was a multicomponent study with separate surveys to obtain information as well as to recruit subjects.

From a large survey to recruit Facebook users (from Craigslist), 19 individuals aged 18 to 56 and with diverse occupations were selected for in-depth interviews. Some of these individuals visited Facebook up to eight times per day. The interviews (1–1½ hours each) used open-ended questions that focused on the use of Facebook, views of privacy, and regrettable experiences of the participants and of their friends. The participants were invited to participate in a diary study in which information on Facebook use would be assessed daily for a month to capture experience and to complement the interview that replied primarily on memory of Facebook experiences. Twelve individuals participated and answered questions daily on a form provided on the Web that related to their Facebook use that day. As it turns out, the interviews were much more useful because participants had few regrets occurred during the 1-month period they recorded their Facebook activity in the diary.

Broad categories not necessarily independent of regrettable topics (e.g., posting sensitive topics about alcohol and substance abuse, sex, religion and politics, profanity and obscenity, personal and family issues), content with strong sentiment (e.g., offensive comments to others, arguments), and lies and secrets (e.g., as jokes or to expose) were identified.

Reasons for regrettable posting too had thematic categories (e.g., it is cool to do it, venting frustrations, effort to be funny, had good intentions, did not think about it very much in advance).

Why do you think the regret was there?

Answers fell into several categories related to unforeseen consequences (unintended audience, underestimated consequences) and misunderstanding how social networking worked (e.g., problems with using Facebook). Finally, another set of categories focused on how to avoid or handle regrets and included trying to keep personal and professional spheres quite separate, delay in sending posts, declining and ignoring requests from others they do know, self-censoring, apologizing, using fake names, having multiple Facebook accounts, and others.

Again the categories are important, but the richness of detail cannot be conveyed. For example, in the set of studies in this report, 574 different regrets were noted with most of these related to sensitive topics as highlighted above. The authors drew on impression management theory that emphasizes how we perform differently in different contexts to help manage impressions. A difficulty with

social networking is that one can unwittingly convey information to an unintended audience (e.g., employers, school admission officers).

9.5.6: Comments on This Illustration

Comment: What have we learned? The richness of the data conveys that regret seemed to be common. While a true prevalence study was not conducted (representative sampling), 23% of the one of the surveys (of 340 users) had regrets about what they sent through Facebook. The topics that served as the basis of regret too are important as reflected in the themes.

The investigators drew on their findings that might be used to influence the design and use of Facebook to help reduce regrets. For example, much of the regretted material was sent in the heat of the moment where strong sentiment and emotional states (e.g., anger, frustration) were involved. There is software readily available that could identify sentiment (emotion)-based messages and postings and the sender might be alerted with one or more "are you sure you want to do this" type messages. More generally, privacy settings and aides on Facebook (and other social networking sites) could convey:

- Common problems
- Sources of regret
- Unintended consequences

On the one hand it seems reasonable to put the responsibility in the hands of the user. Yet, that does not in any way argue against helping to protect both users (from themselves) and potential victims (targets of careless postings that were or were not intended to harm). This study was interesting because it included a back and forth from survey information in the quantitative tradition and then moving to qualitative study of small numbers of individuals in depth. (This combined approach is discussed next.) The main message we might say we already know: Be careful on what you post. Yet, it helps enormously to see the range of likely regrets and how these might be avoided.

9.6: Mixed Methods: Combining Quantitative and Qualitative Research

9.6 Analyze the mixed-methods research that supplements qualitative and quantitative research results

Quantitative and qualitative research methods have different foci and goals when elaborating a phenomenon of interest. It is natural to consider using the methods in combination. The combination has been referred to as *mixed-methods research* and is considered by many to represent a third paradigm of research, adding to quantitative and qualitative research (e.g., Johnson, Onwuegbuzie, & Turner, 2007; Teddlie & Tashakkori, 2012).

Mixed-methods research has its own literature, guidelines, and strategies (e.g., Guest, 2013; Leech & Onwuegbuzie, 2010; Teddlie & Tashakkori, 2012; Venkatesh, Brown, & Bala, 2013).[5] For the purpose of this chapter and text, it is important to convey that combining quantitative and qualitative research methods is not only possible but also an active area of research. Moreover, as quantitative and qualitative methods, mixed methods are used across disciplines and areas of research (e.g., psychology, education, health care, geriatrics).

The larger lesson of this text is that any single methodology (e.g., quantitative, qualitative) or any single facet of a study (e.g., the measures used or methods of data analysis) can limit our understanding of a phenomenon of interest. Using different and diverse approaches can reveal much more than any single one.

In the spirit of this methodological pluralism or methodological diversity, mixed methods provide an important illustration. An example is provided to show that quantitative and qualitative methods are used in the same study and with the benefits of both.

9.6.1: Motorcycle Helmet Use

Motorcycle accidents account for a significant proportion of traffic deaths and injuries. As might be expected, individuals who do not wear helmets have a higher risk of injury and death. Among those who are injured, individuals who were wearing helmets have shorter hospitalizations, lower probability of long-term disability, and less costly medical treatments, as reviewed in the study I illustrate here (Zamani-Alavijeh, Bazargan, Shafiei, & Bazargan-Hejazi, 2011). Motorcycles are particularly popular as a mode of transportation in many Asian and developing countries. Even with mandatory helmet rules, compliance may be low.

This study examined helmet use and factors that were barriers or facilitators of use among motorcyclists in Iran. A mixed-methods approach was used, i.e., both quantitative and qualitative methods in the same report.

The quantitative study began a randomized sampling of roads throughout districts of Tehran, Iran (population of 15 million in greater Tehran and the largest city in the Middle East). At the roads, observers were stationed to code a variety of data from motorcyclists they saw (e.g., passengers, helmet use, of driver and passenger, and type

of helmet—partial or full). A total of 6,010 motorcyclists were observed.

Among the findings, 33% of the motorcyclists wore helmets, another 16% carried them but did not wear them, and 51% did not carry or wear helmets. Of those who wore helmets, only 30% wore the fully protective type. Of the passengers, less than 1% wore helmets. There were some slight seasonal variations. More people (cyclists and passengers) wore helmets in the winter when compared to the summer but that did not alter the conclusions. Helmets were infrequently worn.

> The qualitative study used a sample that was convenient and purposeful to obtain motorcyclists from different circumstances and settings: motorcyclists in the streets, couriers using their motorcycles to deliver mail and packages, those who were receiving medical care, and those who used motorcycles for recreation.

With this group (N = 621) focus groups were conducted with open-end questions (about experiences, how they deal with cars and traffic, how they handle police when pulled over). An additional sample (N = 29) was used for more in-depth interviews and included passengers of motorcycles, police, and family members of cyclists. They were selected because they were readily available rather than through special selection, sampling, or screening procedures.

> The qualitative information was used to identify several barriers and facilitators of helmet wearing. The different themes or types of barriers and facilitators were related to:
>
> 1. Helmet characteristics
> 2. Sociocultural factors
> 3. Personal and psychological factors

A few examples can illustrate some of the findings within these categories. The most frequent impetus facilitating helmet wearing was when there was concern that a traffic officer was going to check. Other facilitators included being encouraged by others (family, spouse) to wear a helmet or being the head of a household with that added responsibility.

The main barrier for not wearing a helmet, especially a full helmet, was its discomfort; also partial helmets were more comfortable. Barriers also included more specific complaints about helmets (e.g., heavy, hot, limit visual and hearing, and "messing up" the rider's appearance, and lack of storage compartment on the cycle). In addition, peer influence was a barrier because in the social context:

- Most others did not wear helmets
- There was a lack of enforcement of helmet laws
- Fear of theft of the helmet

Again with the qualitative results, only a sample of the rich yield is presented here.

9.6.2: Comments on This Example

> **Comment**: Even with the highlighted presentation, one can see the benefits of combining quantitative and qualitative results. The quantitative information conveys the scope of the problem, the conditions under which the problem occurs, and other details that characterize wearing or not wearing helmets. The qualitative information is enormously useful in identifying possible points to intervene that might have impact on the problem. Design and manufacturing of helmets, national campaigns to alter the culture (e.g., use of prominent figures), are merely some of the suggestions that might follow.

It might even be useful to do the next qualitative study to identify all the positive avenues that could be pursued to make helmet use more acceptable, common, and "cool" within the constraint and traditions of a culture. What motorcyclists (or people in general) identify as potentially effective interventions may or may not work but those views are an excellent place to begin. If those views can be complemented by psychological theory that has support in other quarters (e.g., related to adherence in taking one's medicine or in initiating new activities such as exercise) all the better for designing interventions.

I hasten to add that helmet use of motorcyclists is a huge international concern in light of injury and fatalities and has been the subject of other qualitative and quantitative studies. The issue, important in its own right, is part of a larger set of concerns about safety practices where people are not engaging in a particular practice (e.g., bicycle helmets among children, seat belt use among all care passengers, hand washing among physicians between patient exams, not using smartphones or texting while driving). In many cases the practices might be mandated by law and policy, which is not the same as getting people actually to engage in the requisite behaviors. Understanding how to get individuals from knowing (knowledge of what to do and what the law or policy is) to doing (engaging in the requisite behavior) is an enormous challenge. Promoting healthy behaviors has impact on psychological and physical well-being for individuals but also on society at large (e.g., health care costs, emergency room visits).

> Overall the study conveys the complementary information from mixed methods. There is a way in which one wonders why all or most human studies are not mixed methods when the goal is to measure a particular problem or experience of that problem. We need the quantitative data—there is no substitute for prevalence, incidence, risk, outcomes, and the usual from quantitative research. Yet we need to know the qualitative data too because it places meaning on the numbers and guides critical dimensions and domains that are lost in the numbers.

This latter feature is likely to impact on its own because qualitative data bring the quantitative data to light and provide poignant meaning to the numbers. (In fact after reading this, I am wearing a helmet right now as I type and I do not even have a motorcycle.) Quantitative and qualitative research comes from quite different traditions and rarely within psychology training (e.g., graduate school) at least are there opportunities to learn qualitative research, let alone any special strategies that might be added to that for mixed methods.

9.7: Recapitulation and Perspectives on Qualitative Research

9.7 Analyze the pros and cons of qualitative research

It is important to know about qualitative research and mixed methods because of the expanded opportunities they provide for studying a topic beyond the quantitative research methods in which most of us have been exclusively trained. Let me highlight the contributions of qualitative research to recap key points but also to add perspective as you ponder the utility and use in your own research. Also, I have lamented that there are no routine opportunities in most psychology programs to learn mixed methods, but one can seek this out where there are programs or centers (e.g., see articles and editorial board members for the journals mentioned in footnote 5).

9.7.1: Contributions of Qualitative Research

The contribution of qualitative research is its systematic approach to the subject matter. There are formal procedures and guidelines for:

- Collecting information
- Guarding against or minimizing bias and artifact
- Making interpretations
- Checking on these interpretations and on the investigator
- Ensuring their internal consistency and confirmability of the findings
- Seeking triangulation of methods and approaches to ensure that the conclusions are similar when the methods of study are varied
- Encouraging replication, both within a particular data set (by other investigators) and with additional data (e.g., multiple cases)

The purpose of qualitative research is to understand, elaborate meaning, and to uncover the experience of the participants. The multiple ways in which this can be achieved are systematic and replicable, and they include formal procedures for data collection and evaluation. There is not much more we can ask of an empirical approach to psychological phenomena than the formal procedures central to qualitative research.

Qualitative research makes several contributions to knowledge.

> Qualitative research can elaborate the nature of experience and its meaning.

Needless to say, the information and level of analysis does not replace, compete with, or address the yield from quantitative research. As I mentioned, for so many topics (e.g., mental and physical disorders, crime, substance use, and endless more) we need information on the incidence, prevalence, risk factors, and tests of causal models to identify and understand differences between groups (e.g., depressed vs. not depressed) on a variety of other measures (family history, cognitive processes, and so on). This requires information from quantitative studies.

Yet, the information from quantitative studies, however important, omits the richness (depth and level of analyses) of individual experience. Also, the quantitative research quite purposefully often simplifies and omits many variables (for purposes of control) to study some smaller set of variables. Often the task is to arrange experimental conditions to isolate a variable or to study how it interacts with one or two other variables.

> Qualitative research emphasizes many variables in their multiplicity and contexts and brings to bear another level of analysis by elaboration and consideration of the details.

The kind of understanding sought by a qualitative approach seeks the full richness and bloom of some experience or phenomena. The approaches do not compete; they are clearly complementary.

> Qualitative descriptions can serve as an unusually good basis for developing as well as testing theory.

Developing theory as an initial stage warrants further comment because that is its special strength.

> Qualitative researchers often speak of *grounded theory*, a term used to reflect the development of theory from careful and intensive observation and analysis of the phenomenon of interest.

That is through close contact with the details of the phenomenon, analytic interpretations are developed; these are refined through rechecking and further confirmation.

The process begins by obtaining information that reflects participant experience including thoughts, feelings, context, and so on. Once that information is obtained, one seeks to identify abstractions, themes, and categories. There are guidelines, questions to ask of the data, and methods of coding that are designed to abstract these themes that do not make presuppositions of how the experiences ought to be organized. Once tentative themes are identified, there is an iterative process of going back to the original information and again to the themes to check and revise. Clearly, the abstractions (theory) are *grounded* in and close to the data of the participants' experiences (Bryant & Charmaz, 2012; Wertz, Charmaz, & McMullen, 2011). Interpretations and analyses of the data generate hypotheses that can be pursued in further research, whether qualitative or quantitative. By design, qualitative research usually begins with experience as presented without preconceived notions of a theory or the best way to capture and characterize experience. The resulting theory can be tested and further evaluated in quantitative and qualitative research.

> Qualitative research provides an excellent basis for developing one's own research in a given area.

Early in the text I discussed sources of ideas for research. Too often we engage in research and on topics for which we have great interest but also for which we have had relatively little direct contact. For example, one might be interested in suicidal ideation; the experience of trauma; a special patient group; individuals with panic attacks; or terminally ill children, adolescents, and adults with enormous optimism. One can "read the literature" and work on a dataset, but there is no substitute for in-depth exposure to and experience with the group of interest. Participation and engagement with individuals directly will provide a flood of ideas and options about what is going on, why, and what might be done about it. One's own thinking will generate many ideas but so will direct comments of the individuals themselves as one gets to know them and their experiences and stories. What an excellent way to develop a base for theory! Developing and testing theory are a back-and-forth process, so one can also do another qualitative study to test one's theory about experiences and different ways individuals construct them.

One can see the benefits of intense and careful study of individuals outside of the formalities of qualitative research. In clinical psychology, psychiatry, and closely related areas, major advances have been made by studying individuals in depth. Examples from psychiatric diagnosis (e.g., Kraepelin on diagnosis), neuropsychology (e.g., Phineas Gage brain injury), psychotherapy (e.g., Anna O. and other cases of Freud), and learning-based psychological interventions (e.g., Watson's and Jones's work on Peter to overcome anxiety in a child) do not scratch the surface of individual cases that exerted extraordinary impact. These cases have spawned decades of research and understanding at multiple levels. The impact is dramatic (and occasionally baffling) in some of these areas because the original cases seem to have been largely misinterpreted (Anna O looks like she did not get better at all or not from "talk therapy" and talk was only one of the interventions) or could not be replicated (e.g., Watson's with Little Albert). Yet the point is that knowing a phenomenon in-depth permits one to generate hypotheses about what the key constructs are for understanding that phenomenon and what the likely causal paths and influences are. As important, exposure to the individuals of interest can help dispel stereotypes and preconceived notions that often are the initial obstacles to creative research. Qualitative studies add to this by introducing science to enrich the experience by bringing out information is systematic, empirical, and replicable ways.

> Qualitative research can suggest causal relations and paths over the course of development.

Quantitative approaches have made enormous gains in identifying multiple factors and their contribution to a particular outcome. The findings are valid at the level of group analyses. It is important to know the general variables that are likely to yield a particular outcome, but also to view these in the contexts in which they may or may not operate. The causal sequence and path leading to an outcome for *individuals* is not really addressed in quantitative research, although some lines are moving in this direction (e.g., personalized medicine). The arts and literature can provide intriguing insights and generate many hypotheses about the individual, but a more systematic, empirically based approach to elaborating the richness of individual experience is needed.

> Qualitative analyses provide a systematic way of looking at potential causal paths, unfolding of events, and dynamic and reciprocal influences of events for individuals (see Maxwell, 2012).

These can be tested experimentally in quantitative studies with statistical evaluation and perhaps by direct intervention.

> Qualitative research looks at phenomena in ways that are intended to reveal many of those facets of human experience that the quantitative tradition has been designed to circumvent—the human experience, subjective views, and how people represent (perceive, feel), and hence react to their situations in context.

For example, quantitative research has elaborated many of the factors that contribute to or are associated

with homelessness (e.g., crime, physical and mental illness, hunger, victimization). A qualitative study is likely to focus on the experience of being homeless, the details of the frustrations, and conflicts and demands the experience raises in ways that are not captured by quantitative studies (e.g., Nettleton, Neale, & Stevenson, 2012; Stevenson, 2013).

Similarly, quantitative research has elaborated the worldwide epidemic of HIV/AIDS, developed effective treatment and preventive interventions, and more. Qualitative research can describe in detail what life is like on a daily and indeed moment-to-moment basis to learn of one's diagnosis, to interact with one's partner and relatives, to worry about taking medication, and to face death that make the experience of HIV and AIDS vivid and poignant (e.g., Kempf et al., 2010; Vaz, Eng, Maman, Tshikandu, & Behets, 2010). Such analyses can very much move others and have remarkable impact; it can also inform research about how to better attend to the care of those who suffer HIV/AIDS or are caring for and living with them. An anecdotal case study might do this. Yet, we need systematic understanding that can build a knowledge base. That requires strong methodological tenets and practices, and the scientific approach of qualitative research provides that.

9.7.2: Further Considerations Regarding Contributions of Qualitative Research

Although qualitative and quantitative research derive from and pursue somewhat separate traditions, they can be combined in various ways.

As highlighted and illustrated in the comments on mixed methods, increasingly qualitative and quantitative methods can be combined in the same study. For example, one can use the rich detail and extensive records of the qualitative analysis for testing as well as for generating hypotheses. Coding the content and looking for themes and sequences not only describe interactions but also pose and test the extent to which some events or explanations are plausible. The constructs and categories that emerge from qualitative analyses can be used to develop new measures, i.e., new ways of operationalizing concepts for empirical quantitative research. Indeed, measures would probably be much better in capturing constructs of interest if they began from in-depth appreciation of the construct and how individuals experience life. When we develop a measure, there is often a concern with the psychometric properties, i.e., many forms of reliability and validity. Qualitative analysis in this context alerts us to other issues, namely, the extent to which experience is suitably captured by the items and relevance.

The qualitative research can contribute greatly to our understanding of cultural, ethnic, and groups of varying identity.

Quantitative research may analyze differences in a study and show that culture or ethnic group serves as a moderator. That rarely increases our understanding. In-depth evaluation of how different cultures experience the world and the phenomena we study could help provide a truly culturally sensitive social science.

To give a concrete example, many evidence-based interventions have been developed in psychology. Occasionally but not too often do we study whether different cultural groups profit from the treatments or experience the treatment differently. For example, parent management training is a very well-established treatment for children who engage in oppositional, aggressive, and antisocial behavior (Weisz & Kazdin, 2010). A recent qualitative study of Latina mothers revealed that components of the treatment vary in the extent to which they are perceived as relevant or acceptable (Calzada, Basil, & Fzernandez, 2012). In adapting treatment to cultural issues, such information could make a huge difference regarding seeking treatment, remaining in treatment, and adhering to procedures once someone is in treatment. Here is a case where drawing on qualitative research findings might readily enhance the use of findings obtained from quantitative research.

In the context of treatment, there is another where qualitative research or mixed methods are sorely needed. In psychosocial treatment evaluation, researchers examine whether the changes clients have made at the end of treatment are clinically significant, i.e., a change that has genuinely affected their lives. Sophisticated and fancy indices of change on psychological measures, blessed with many psychometric properties, have been devised. The difficulty is that no matter how a client scores on most of the measures, we still have no idea about whether the change affects the daily life of the client or whether he or she is better in any palpable way (Blanton & Jaccard, 2006; Kazdin, 2006). An extreme aberration of clinical "irrelevance" may be concluding that treatment really helped based on "effect size," a metric discussed in detail already. Effect size is so very far removed from patient experience and any sign that patients were helped in palpable ways (Kazdin, 2013b).

Developing the measures through qualitative studies would be one strategy to identify relevant outcome assessment. That is, measurement of an important and clinically significant change could begin with an intensive evaluation of client experience after treatment among those who feel as having changed or among those who are so identified by their relatives. Developing measures

well-grounded in the experience of clients would be valuable in their own right. In addition, those measures could be subjected to usual methods of scale development and evaluation and lead to systematic research on clinical significance.

> Finally, delineating the unique features of qualitative research also helps one understand the strengths, contributions, and limitations of quantitative research.

By way of analogy, the purpose of requiring undergraduate college students to learn a foreign language is not only to expose them to many facets of another culture but also to bring to light and to provide perspective on their own language and culture.

Learning another language brings to light features of one's native tongue, by making the tacit explicit, and by seeing how similar goals (e.g., conjugating verbs, using gender based pronouns, expressing joy or disgust) can be achieved in different ways and with slightly different emphases. Discussion of qualitative research has parallel benefits. We take as a given that quantitative research methods (hypothesis testing, statistics) are the only way to obtain empirical knowledge and if not the only way then certainly the best way. There is no need to take either of these positions. Qualitative research is different, but it is research and empirical and can be evaluated on its own grounds. Highlighting qualitative research can bring into sharp focus critical issues of quantitative research that serve as the central basis of this text. One of the lessons of methodology is that how a phenomenon is studied (e.g., what designs, what measures) can directly influence the results. This lesson fosters humility about any one method and encourages use of diverse methods within any given field or topic.

For all of these reasons, qualitative research can play and arguably ought to play a stronger role in research. Qualitative research has many different methods and is used in many fields of study. That can lead to somewhat different meanings in how the term is legitimately used. Yet let me end with a cautionary note. Often the term is misused in clinical psychology, as well as in other disciplines (e.g., psychiatry, sociology). Qualitative is sometimes used to refer to descriptive, anecdotal, and case study material. That is, the term has been inappropriately adopted to refer to any nonquantitative evaluation. This is a misuse—qualitative is not a synonym for loose, unsystematic, or "my opinions" or for "I don't really collect data." Be wary when you hear the term casually tossed. Qualitative research is:

- Rigorous
- Scientific
- Disciplined
- Replicable

9.7.3: Limitations and Unfamiliar Characteristics

There are characteristics of qualitative research that are unfamiliar and worthy of note. Some of the characteristics would be limitations in quantitative research but are not in the context of qualitative research:

1. Unfamiliar Characteristics 1 and 2 of Qualitative Research: Qualitative research generally does not use a control group.
2. Unfamiliar Characteristics 1 and 2 of Qualitative Research: Qualitative research relies heavily on self-report.
3. Unfamiliar Characteristics 3, 4, and 5 of Qualitative Research: Qualitative research depend very much on the investigator in devising a study.
4. Unfamiliar Characteristics 3, 4, and 5 of Qualitative Research: Assessment of experience in much of qualitative research does not consider a developmental, life-course perspective. That is, one "context" for experience has to do with a particular point in time and one's constructions can change greatly.
5. Unfamiliar Characteristics 3, 4, and 5 of Qualitative Research: Ethical issues and participant protections pervade all research (e.g., informed consent, protection of privacy, and many others).

Each of these characteristics are discussed in detail in the following sections.

9.7.4: Unfamiliar Characteristics 1 and 2 of Qualitative Research

> First, as you have noted from my examples qualitative research generally does not use a control group.

The task of research is to elaborate how a particular group of interest experiences some facet of life and a control group does not have the same role as it would in quantitative studies.

Conceivably, there might be use of a control group. For example, in the study of bus crash victim, the focus was on their reactions. It would make no sense to ask a control group (e.g., individuals who just finished a bus ride with no crash) how they felt about that.

What would one want to control for?

A possible reply to this might be including a group who experienced personal injury that is not part of a larger scale disaster (e.g., individuals seen at a hospital for broken bones or injury that requires intervention). Maybe the themes and issues that emerged with bus crash victims would emerge among any individuals going through injury. It would be useful to know what larger scale disaster experiences are unique compared

to the more common and routine personal injuries and crises we experience. My comments are a stretch and reflect quantitative research thinking; the goal of the qualitative research study was to elaborate the experience of one group. My "control" idea raises the question of whether the type of injury/accident makes a difference, i.e., moderates the relations.

> Second, and as a methodological feature, qualitative research relies heavily on self-report.

Obviously, when one wishes to understand experience, we need humans to report on that experience. That is why in-depth interviews, audiotaping, and then transcribing the interviews are completed.

Yet, the limits of self-report are important to keep in mind. For example, in one of the examples (regret in Facebook postings) individuals were asked to identify why they did something. Yet people are not necessarily in a position to comment on why they did something, although our views and stories might be interesting in their own right.

We all have stories and firm views about what influenced us here or there, but there is no evidence that the factors we report as influential really account for the paths we have chosen and in fact often competing evidence of strong or multiple factors we neglected to mention (e.g., genetics, temperament, priming cues in the environment) that were causally involved. There may be multiple influences, many of which we cannot verbalize, and these influences vary in the weight (strength of impact they exert). We are not in a position to identify in fact the scope of influences and their weight, by the virtue of human limitations (e.g., cognitive heuristics, implicit attitudes, impact of learning experiences at different stages, temperamental influences). This is a limitation of self-report and not qualitative research. I mention it here because qualitative analyses with humans rely so heavily on self-report.

The counterargument to the concern about self-report is that perceptions and verbal statements in qualitative research are important in their own right and not because they map on to some other index that is a criterion to evaluate **"accuracy."** One's construction of an event or experience is essential and valuable because that is one's experience. The value does not come from correlation with some other metric. Another counterargument will have special experience to uncompromising quantitative researchers. Subjective experience and perception do map on to many other important **"objective"** indices. For example, subjective experience (e.g., of stress, happiness, loneliness) influences our immune systems and gene expression in an ongoing way that affects physical and mental health and longevity (e.g., Hawkley & Cacioppo, 2010; Slavich & Cole, 2013). That nasty outcome measure (e.g., early death) is in here. How we all experience the world is important whether one takes a qualitative or quantitative research perspective.

9.7.5: Unfamiliar Characteristics 3, 4, and 5 of Qualitative Research

> Third, the qualitative research depends very much on the investigator in devising a study.

Open-ended questions are usually provided to allow participants free reign in constructing experience and stating what is important, why, when, and how. All that is to the good. Yet, we now know that humans bring to situations cognitive heuristics of all sorts that greatly frame how things look and the conclusions we reach. This is not a "limitation" but just a characteristic very much like the characteristics of human vision that see a very small part of the spectrum of light and color. These same cognitive heuristics are in the repertoires of participants. What all this means is that experience as evaluated and constructed is heavily filtered by all participants (investigators and subjects). Much cannot be put into words and what may not reflect tacit and ineffable facets of the experiences. Much has been blocked out of recall or was not perceived to begin with. These are givens among humans as we observe, record, recall, and organize experience. It is not clear to what extent these considerations have played a role in evaluating qualitative research. Some of the critical principles and procedures (e.g., trustworthiness, confirmability) do not necessarily address these matters.

> Fourth, assessment of experience in much of qualitative research does not consider a developmental, life-course perspective.

That is, one "context" for experience has to do with a particular point in time and one's constructions can change greatly. For example, we experience something (e.g., relationship) this way now, that way then, and that new way later.

> Experience of a phenomenon is not at all fixed and is better conceived as a video rather than a still photo.

For many of us, our first love in a romantic relationship at that point in time was often experienced as the only love, the only person, and the only soul mate—what a heavenly match—there is no other person for us. That could change 1 year into the relationship and 10 years after the relationship is ended, or much later—perhaps even three marriages later (what was that person's name again—it is on the tip of my tongue). Soul (sole) mate can switch to heel mate sometimes. All of these are experiences; all are veridical, but they may differ greatly for an individual over time.

Some qualitative research is conducted longitudinally over a period of few to several years (see Neale, Henwood, & Holland, 2012). For example, one study characterized the difficulty female medical students experience and how they respond to sexual harassment and gender discrimination

(Babaria, Abedin, Berg, & Nunez-Smith, 2012). Another qualitative longitudinal study tracked individuals with chronic health conditions and what they came to know and used as information to manage their condition (Edwards, Wood, Davies, & Edwards, 2013). Clearly longitudinal work is available, but so many of key questions about experience in psychology change over one's life course. Much more attention is needed in charting experience. More qualitative research is needed with a longitudinal and prospective perspective. This research is challenging because methods of capturing experience (e.g., from toddlers through the elderly) vary in the challenges they present.

> Finally, ethical issues and participant protections pervade all research (e.g., informed consent, protection of privacy, and many others).

Qualitative research can raise another issue of which researchers are well aware. If individuals are asked to reflect on their experiences in depth, this might have its own untoward side effects. For example, a qualitative study of victims of domestic violence or terrorism might require in-depth recounting and re-experiencing events. The thoughts, emotions, and cognitions might well make the person "worse" in some temporary or enduring way. That is, a person has come to grips, learned to regulate, coped with something, or has successfully sequestered that past from all current experiences and thoughts. Now in a qualitative research project, the experience is rekindled, fueled, and burns freely. This does not necessarily lead to problems but is an ethical sensitivity that is required. The issue can be seen in the context of quantitative research. Often in quantitative research, review boards at the university are deeply concerned about an item on a scale (e.g., of depression) if that item mentions suicide or suicidal ideation. Occasionally, the investigator will be asked to delete that item. Consider this sensitivity. Now return to qualitative research that may well focus in great depth on the topic. Balancing scientific yield and potential cost and risk can be more challenging in this context, depending on the topic of study.

9.7.6: General Comments

The contribution of qualitative research is easy to argue because of the importance of diverse methods in general as well as the specific and unique contributions in understanding experience. Moreover, the ability to combine quantitative and qualitative research provides opportunities for novel elaboration of phenomena. There remain obstacles in advancing qualitative research.

> Arguably the greatest obstacle is the absence of training opportunities. Qualitative methods are infrequently taught or taught in sufficient detail to be able to carry out a study for most individuals in training.

Among the reasons, researchers must learn the methods in the quantitative tradition if they are to enter the research system in academia, at least in the United States. Also, quantitative research methods continue to evolve with:

- Novel experimental designs
- Measures (e.g., continued breakthroughs in neuroimaging and genetics—increasingly used in psychology and clinical psychology)
- Methods of data analyses (e.g., to capture networks, systems, and dynamic states) (e.g., Little, 2013)

Indeed, many fairly old methods within the quantitative tradition (e.g., Bayesian analyses) are taught infrequently but enjoy a new resurgence. In other words, training and equipping future researchers or those who read and consume research with current and evolving tools in the quantitative tradition is a challenge. Now consider adding to that additional methodologies (e.g., single-case, qualitative, mixed methods). There are realistic limits to what training is likely to provide with many extra experiences (e.g., summer workshops, postdoctoral positions).

An important advance in quantitative research has been the development of meta-analyses as a method of combining multiple studies.

> Meta-analysis is now very familiar and is routinely used to combine the results from research in a quantitative tradition. Studies from qualitative research often are neglected. Is there a way to bring many such studies together in a way that is parallel to meta-analysis?

Actually, there are ways to accumulate and integrate findings from qualitative research. These include a set of procedures referred to as *meta-synthesis* (Sandelowski, 2012). Research syntheses of research are published primarily in journals where the methodology is respected and common. This has tended to sequester the accumulated knowledge from researchers in the quantitative tradition. A similar situation occurred for single-case designs that for many years were restricted to those journals that only published single-case experimental designs. It is still the case that in summarizing research on effective interventions, findings from single-case designs are neglected. Among the reasons have been the lack of agreed-upon ways to evaluate effect sizes and incorporate results in meta-analyses of group quantitative research (see Kazdin, 2011). The challenge for our science is to bring to bear all of the methodological approaches that can elaborate a given topic or problem. Quantitative (between-group, single-case), qualitative, and mixed methods are broad categories to convey multiple approaches. This is beyond the purposes of the text, but it is important to mention that we pay a price in not attending to the full range of scientific methods available to us.

Summary and Conclusions: Qualitative Research Methods

Qualitative research is designed to:

- Describe, interpret, and understand human experience
- Elaborate the meaning that this experience has to the participants

The data are primarily words and are derived from in-depth analysis of individuals and their experiences. A key feature of the approach is a detailed description without presupposing specific measures, categories, or a narrow range of constructs to begin with.

The approach differs from the dominant research paradigm, referred to as quantitative research, in how the study is completed, the roles of the investigator and participants, what the data are, how they are examined, and the nature of the conclusions. Although extensive data (e.g., narratives, case descriptions, video or audio records) are collected, usually they are not reduced in a quantitative way. Rather, from the data, interpretations, overarching constructs, and theory are generated to better explain and understand how the participants experience the phenomenon of interest.

There are major differences in qualitative and quantitative approaches, but there are also fundamental similarities that make them both empirical research. Among the key similarities are:

- Interest in reliability and validity of the methods of procuring the data
- Efforts to address sources of bias that can impede the conclusions that are drawn
- Replication of both how the study was done and the conclusions that are reached
- Accumulation of knowledge verifiable by others

Qualitative research can contribute to psychology by elaborating the nature of experience and its meaning by bringing everyday but also rare experiences into sharp focus.

Mixed-methods mentioned consist of the combination of quantitative and qualitative research within the same study. Among the obvious benefits is bringing to bear multiple levels of understanding of a phenomenon of interest. Also, the methods allow opportunities for each part of the investigation to influence the other (e.g., qualitative findings can be used to develop measures that will be used in quantitative research). It is important to be familiar with qualitative and mixed methods because of the rich opportunities they provide and because these methodological approaches are much less frequently taught and used in psychological research in comparison to quantitative methods. Moreover, knowledge of these methods places in perspective quantitative methods and their very special strengths as well as their limitations.

Critical Thinking Questions

1. What are the unique characteristics of qualitative research when compared to quantitative research?
2. What makes qualitative research scientific? What practices make this research rather than just anecdotal reports?
3. Make up an example of a mixed-methods research study and what complementary information the study might yield.

Chapter 9 Quiz: Qualitative Research Methods

Chapter 10
Selecting Measures for Research

Learning Objectives

10.1 Examine some of the key considerations when selecting a research measure

10.2 Examine the three avenues of choosing the appropriate measure in research

10.3 Report the need to be cognizant of related issues while choosing the applicable measures

10.4 Describe the implications of the terms brief measures, shortened forms, and use of single-item measures

10.5 Identify three explanations as to why the results obtained through multiple measures may vary

10.6 Examine convergent and discriminant validity

10.7 Review the need to ensure that the selected measure assesses the construct of interest

Systematic assessment is a fundamental across all of the sciences and accounts for enormous advances. Consider some familiar and unfamiliar examples in natural, biological, and social sciences:

- **In astronomy**, our earthbound and orbiting telescopes, more powerful than ever across the full wavelength spectrum, identify objects that are mysterious (never identified before) and less mysterious but new to us such as planets outside our solar system that appear to have the conditions close to those on earth and therefore in principle could be habitable. Habitable means that life might be on the planet and if so a new place for firmly establishing all the familiar fast-food restaurants.
- **In neuroscience**, neuroimaging of human and nonhuman animals has opened new ways of looking at the brain and generating and testing hypotheses about what process might be involved in learning, memory, cognition, social behavior, decision making, overeating, gambling, and more (e.g., Perkel, 2013; Sporns, 2010). The now very familiar methods (fMRI, PET) are early in the assessment methods and much finer-grained assessments of networks, circuits, and individual cell functioning are not on the horizon—many are here that can examine processes at the cellular, molecular, and now atomic level.
- **In genetics**, all sorts of assessment breakthroughs are evident and presented in the news as efforts are made to recover DNA and to have the possibility of cloning extinct species (e.g., Wooly mammoths, various dinosaurs, and Neanderthal people). More evident in everyday life is genetic mapping. In some applications, genetic testing can identify whether some individuals are at high risk for various diseases including various forms of cancer and heart disease but also forms of mental illness.
- **In anthropology**, tombs, lost civilizations, and buried cities and structures can be identified from space satellites (e.g., Schuetter et al., 2013). Once identified from space, the leads that could not be otherwise observed can be pursued directly on the ground to unearth previously hidden structures.
- **In ethology** or the study of nonhuman animal behavior in natural settings, we have learned about migration patterns, diet, social structure, and so much more through advances in assessment. Consider a recent example regarding our knowledge of cheetahs (I mean the animals and not individuals who lie on their income tax). Cheetahs are great hunters not merely or primarily because they are fast as we have always thought (Wilson et al., 2013). Outside of TV commercials, they do not use their great speed most of the time

when they hunt. Rather their hunting prowess comes from unusual agility in starting, stopping, and turning on a dime—more like a nickel.

- **In psychology**, fundamental advances in the psychology of learning, as reflected in the work of Ivan Pavlov (1849–1936) and B.F. Skinner (1904–1990), derived from novel and creative conceptualizations of the subject matter. Yet novel assessment methods used in the research (e.g., drops of saliva, rate of response) contributed greatly to the advances. The measures provided very precise objective measures that were responsive to training and a variety of experimental manipulations.

In each of the above examples, the news media understandably report the interesting tidbits of *what* we have learned. Yet in the background and rarely presented to the public at least is that the findings reflect breakthroughs in assessment.

The assessment methods allow more precise ways of revealing the subject matter of interest to the diverse sciences and often in the process reveal "new" phenomena—they were always there—we can now detect or see them or that look at "old" phenomena but at a new finer grained level of analysis.

All of this is to convey that assessment plays and has played a central, even if often unsung role, in scientific advances (see Greenwald, 2012). These advances are evident in psychological science and clinical psychology as a subarea within that. For example, neuroimaging and genetics have elaborated key topics of interest (e.g., aggression, violence, symptoms of depression, commonalities shared many psychiatric disorders, and changes when psychotherapy is effective).

Assessment in the context of our own studies has a much narrower and more concrete focus. We are going to test a conceptual view or hypothesis (e.g., about emotion regulation, empathy). Among the decisions for the study is selecting or developing the measures we will use.

This chapter focuses on overarching issues that influence or ought to influence measurement selection for research. We will cover validity and reliability of measures and how they influence measurement selection, the use of available measures and the development of measures when those are not too helpful, and many special issues and considerations such as the use of short forms, assessment biases, and others. Assessment is so critical to research and often is an afterthought or ancillary part. The two chapters identify key issues that can guide measurement selection.[1]

When we consider identifying measures for research, attention usually is drawn to the dependent measures or outcomes that will serve to evaluate the hypotheses. Yet, assessment entails all of the measures and their different purposes before, during, and at the end of the investigation:

- Assessments at the beginning of the study *before* the experimental manipulation may be used to evaluate demographic variables, to assess subject characteristics (e.g., exposure to abuse, self-control) that may relate to the outcome (i.e., moderators) to invoke subject selection criteria (inclusion and exclusion criteria), and to provide a base (e.g., pretest) of performance to pre- and post-measures to evaluate change.
- Assessment *during* an investigation (e.g., middle) might be used to check to see if participants really experienced the experimental manipulation or to assess intervening or mediating processes (e.g., emotion regulation strategies, alliance in therapy, cognitions or motivational states, and neurobiological processes [i.e., mediators and mechanisms]) that might explain or correlate with the changes over time.
- Assessment at the *end* of the investigation includes the dependent measures, i.e., those measures expected to reflect change. Also, at this point participants may be asked to reflect on features of the study, what they remember, know, and so on. Sometimes whether they experienced the experimental manipulation to which they were exposed is assessed at this time.

I have highlighted three temporal points (pre, mid, post) where assessments are routinely used, but of course this is a guide with exceptions you already know.

For example, longitudinal investigations assess participants over years and decades. Also, single-case designs assess behavior continually throughout the investigation. Other distinctions of when and how assessments are administered could be made as well.

But the key point is that assessment is central to different facets of the study at different points in time and with many different purposes. Measurement selection is critical.

Assessment is so critical to any study that it is important to provide guidelines and tools for selecting and perhaps avoiding certain measures. The emphasis will be on the dependent measures used for a study, but the key points pertain to assessment and selection of measures more generally.

10.1: Key Considerations in Selecting Measures

10.1 Examine some of the key considerations when selecting a research measure

The selection of measures for research is based on several considerations related to the construct validity of the measure, psychometric properties, and sensitivity of the measure to

the changes or differences predicted by the hypotheses. In a sense, we are "auditioning" possible measures for use in our study. It is not enough that a given measure appeared in a prior study, although that could be pivotal (e.g., for replication of one's own or prior work). We want to invoke several considerations for including the measure.

10.1.1: Construct Validity

As a general rule, in our research usually we are not interested in measures. An exception of course is when we are developing or evaluating a new scale or measure or making a career out of a particular measure that we believe is important and want the rest of the world to use. In the usual case, we are interested in *constructs* or the concepts that these measures reflect. As obvious as this sounds, it is critical to bear in mind, because measures often are selected without sufficient attention to the extent to which they actually measure the construct or facet of the construct of interest to us. For example, there are many measures of stress or social support, two constructs often studied in clinical, counseling, and health research. But the measures are not interchangeable, do not invariably examine the same aspects of stress or social support, and may be quite different in their utility and relevance based on the facets of the constructs they emphasize. Also, investigators occasionally make up a couple of items for their study that are considered to measure support and provide no data to even suggest support of what is being measured.

Consider an example. Social support, stress, and clinical dysfunction are topics that encompass many areas of research (e.g., diagnoses, cross-cultural studies) in clinical psychology (e.g., Maulik, Eaton, & Bradshaw, 2010). At a clinic where I work, parents of aggressive and antisocial children complete a brief measure of social support that has been in use for some time (Aneshensel & Stone, 1982). The measure includes three scales (items):

- How often they have received support from someone (in the past 2 months)?
- How many supportive relatives are in their lives?
- How many supportive friends are in their lives?

The theory underlying the measure is that in times of stress, individuals seek social support and that support can buffer (mute, mitigate) stress. For purposes of my comments here, consider the three scales that all measure social support. In our work, the inter-correlations of the scales range from r (N = 300) = .22 to .36—all significant (p<. 001) but all rather small.[2] The overlap (that they are correlated) suggests they measure a similar construct, but the fact that they are so low means they are not interchangeable and findings obtained for one of the scales may not be obtained for another. Of course, there are many more social support scales in the world and bringing them all into the discussion, if data were available, would underscore the point. That is, all measures of a construct may not be highly related and, as the example noted previously, all items within a measure of a construct may not be highly related. That is not only or merely an assessment problem—it is also related to how constructs behave and how we as humans behave.

Let us consider a construct: love of one's significant other or partner. We may truly love our partner but of the 10 (let us say 10 just for discussion sake) indices or behaviors, thoughts, and feelings that are measures of true love, a given person may be only good at 4 of them and mediocre at 3 others, and would not recognize the remaining 3 if they splashed on his or her windshield while driving on a sunny day. In other words, one can be truly in love but not be high (or score highly) on all of the measures (indices). Like what we know from everyday life, one can be athletic or a great athlete without being good at all sports. What this means for research. Look carefully at a measure you are considering. Does it "get at" or capture those facets of the construct you have in mind, i.e., in relation to hypothesis? Are there other measures to be added to the assessment to more thoroughly represent the construct?

As a place to begin, the initial criterion for selecting a measure is evidence that the measure assesses the construct of interest.

> In assessment, the term *construct validity is used to refer generally to the extent to which the measure assesses the domain, trait, or characteristic of interest* (Cronbach & Meehl, 1955).

10.1.2: More Information on Construct Validity

Construct validity has been used throughout this text to refer to a type of experimental validity that relates to the interpretation of the basis for the effect of the experimental manipulation. In the context of assessment, the interpretation of the measure is at issue, namely, to what extent does the construct underlying the measure serve as the basis for interpretation of the measure?

> In assessment, *construct validity refers to the link between the concept behind the measure and research that attests to the utility of the construct in explaining the findings*.

Stated another way, construct validity refers broadly to the pattern of findings and their integration that bears on the interpretation of the measure.

> Construct validity does not reduce to a correlation between measures or the measure and some other criterion. Rather, it involves the accumulation of evidence from diverse sources.

And that can include many different types of validity, as presented later. In addition, how does the measure relate to other constructs and is that in keeping or consistent with what one might expect from the construct we are measuring?

What do you think?

Fundamentally, construct validity relates to one's conceptual model of the measure, what it assesses, how the measure reflects the underlying construct, and how that relates to other domains of functioning (MacKenzie, Podsakoff, & Podsakoff, 2011). Construct validity involves a broad theory of how a concept (e.g., empathy, kindness, volatile personality) "behaves" and therefore what the measure ought to show in many different contexts beyond any single study.

In a given study, the investigator may be interested in measuring "stress" or "emotional distress." There should be some initial assurance that the measure actually reflects the construct. The assurance comes from accumulated evidence that findings are consistent with this construct and findings that another construct that might be related is not very plausible. In other words, we consider a mini-theory about the construct and measure. If I have a measure of stress, individuals with these characteristics or in these situations (e.g., immediately after a national disaster, during final exam week, learning of a disastrous diagnosis) ought to score more highly than individuals not undergoing any of these. This line of thinking continues—stress as a construct would be manifest in what ways? Construct validity is enhanced with supportive evidence showing that these predictions hold up in multiple contexts and that "stress" is a parsimonious and plausible explanation of what is actually measured by the scale.

10.1.3: Reasons for Carefully Selecting Measures

From a more practical standpoint, when one selects measures for a research project, it is useful to go beyond merely saying, "Well one or two other studies used this measure so it must be fine." Skepticism but not cynicism is a key to methodology. Ask, "What exactly has been shown to relate to the measure? What is the evidence that this measure I am considering captures or encompasses the characteristic I wish to assess?" It is easy to be enticed by many available measures into the assumption that they assess a particular construct. Be strong because it is easy to be tricked in selecting measures and thinking you have what you want for your study. Likewise, when reading the results of the study, look closely at the measures that were used. Here are some reasons why.

First, measures usually have names that reflect the construct the investigator *intended* the scale to measure (and, of course, often the name of the investigator as well). Unfamiliar (and fictitious) examples readily convey the sorts of measures that are available such as the Lipshitz Depression Inventory, Stop-Following-Me Scale of Paranoia, or the Kazdin I'll-Bet-You-Won't Jump Measure of Risk-Taking. The names of various measures may be based on supporting evidence that a particular characteristic or construct in fact is assessed (i.e., construct validity) or merely reflect what the originator of the measure had in mind without the requisite evidence.

> Construct validity requires the accumulation of studies to establish what the measure assesses and whether performance on the measure across diverse circumstances would be expected in light of the construct.

Second, one may read the items and nod in agreement that they in fact are measuring what you want to measure. We will return to this notion of face validity, i.e., whether a measure looks on the face of it that the items are related to the construct. Yet, what items measure can differ from what they seem to measure. For example, in clinical psychology going back decades there have been measures of psychopathology, i.e., how many symptoms of psychiatric disorder or psychological dysfunction one has and how severe they are. Items seem obvious as people are asked to endorse whether and to what extent they are anxious, depressed, hear voices, and so on. Yet construct validity was not so clear because another construct often seemed to be involved (confounded with) called *social desirability*, i.e., placing oneself in a favorable light. That is low symptom scores might reflect few symptoms but also could reflect a response set of trying to look good or better than one really is by not endorsing symptoms. (I will mention social desirability in more detail later because it can play a major role in assessment.) In short, of course look at the items, but do not assume that the content as described in fact is what is really measured by the item(s) or would relate to other well-established measures of the construct the items appear to measure.

Finally, many questionnaires and inventories have been evaluated through factor analysis, a set of statistical procedures designed to identify correlated items that cluster together. Names of the factors also suggest what is "really" measured.

Essentially, the factors (subsets of items) may be presented to assess different constructs or different facets of a single construct. Here too, one must be cautious because the connection between the name of the factor and what the items have been shown to measure (construct validity) is not always clear. Also, whether a scale, factor, or subscale with a given name is what the investigator means by the construct underlying the investigation is not automatic.

For example, one might develop a measure of love, but there are different types of love and different relations in which they are manifest. Is this measure of love the type the investigator has in mind, or is it distinguished from other types of love, or from liking a lot, from loyalty, from positive affect ("warm fuzzies") in general that is not necessarily love? As a construct validity exercise, ask questions the next time someone says, "I love you." Ask the person, "What

kind of love? What evidence can you offer that you do not just mean, 'like me a lot?' Maybe it is just physical attraction because I am drop-dead gorgeous?" (Methodologists always ask a lot of questions like these which is why most of their relationships are statistical rather than interpersonal.) These are construct validity questions and are not trivial. Needless to say, there should be some evidence that the measure selected for research in fact assesses the construct of interest. If there is no evidence available, some steps within the study should be taken to provide information on validity and other psychometric properties.

As a summary guide to construct validity, it is fair for one's advisor, mentor, or colleague to ask the question as you design your study, "what makes you think those measures assess the constructs you care about?" Similarly, if you are reading an article and the title suggests constructs (e.g., violence, aggression, kindness, moodiness) be sure to look closely at the measure(s) to see what was actually assessed how strong the evidence is supporting the connection between the measure and the construct. Construct validity is a guiding conceptual answer—the evidence from multiple studies that the construct behaves as expected and the measure is not likely to be explained by other constructs. How does one really identify if these latter criteria are met? The answer comes from the current evidence on the psychometric characteristics or properties about the measure, which coincidentally is the next topic.

10.1.4: Psychometric Characteristics

There are many steps for establishing or deciding whether a measure adequately assesses the construct of interest. These steps, broadly conceived, refer to how the performance on the measure relates to other domains of functioning (other measures) across a variety of circumstances. There are several characteristics of performance on the measure that contribute to the construct validity of the measure.

> *Psychometric characteristics is a broad term that refer to the reliability and validity evidence in behalf of a measure*.
>
> *Reliability generally refers to consistency of the scores obtained for the measure, including consistency among items of the measure (i.e., how the items relate to each other), consistency between different parts or alternate forms of the same measure, and consistency in performance on the measure over time (test–retest for a given group of subjects*).
>
> *Validity refers to the content and whether the scores on the measure are shown to assess the domain of interest and encompasses the relation of performance on the measure to performance on other measures at the same time or in the future and to other criteria with which the scores on the measure would be expected to be related (e.g., school achievement, occupational status, psychiatric diagnosis).*

Any single definition of reliability and validity is hazardous because both are broad concepts and each has several subtypes. Also, over the years the different types of reliability and validity and their meanings and terminology have varied, so there are some inconsistencies (slight unreliability) in use of the terms reliability and validity(see Wasserman & Bracken, 2013; Watson, 2012).

10.1.5: More Information on Psychometric Characteristics

Table 10.1 presents major types of reliability and validity that are commonly referred to and of clear relevance in evaluating whether or the extent to which a measure is suitable for one's work or project. The terms in the table are useful to know because they are core concepts within assessment. Also, the concepts of reliability and validity sensitize the investigator to a range of considerations. In any given situation, a specific type of reliability and validity may or may not be relevant.

For example, high *test–retest reliability* (e.g., Pearson product moment correlation or *r* that is > .7 or .8) over a period of a few months might be expected for a measure designed to assess a stable characteristic (e.g., traits such as self-control, extroversion, or altruism) but not for more transient characteristic (e.g., a state such as irritated mood or perhaps anger). Similarly, a high *concurrent validity* correlation of the measure (e.g., of social support) with other indices of the same construct (e.g., frequent activities with friends, reliance on friends for help) would also provide support for the measure. Yet, a high correlation of that same measure (of social support) with indices of other constructs (e.g., intelligence, conscientiousness) might raise problems. This is why construct validity relates to "theory," i.e., one's conceptualization of how the construct relates to other constructs and also draws on multiple studies. Construct validity is not about one or two correlations but rather if performance on the measure across different situations "behaves" as one might be expected by the concept underlying the measure.

> Measures known to reflect the construct of interest, and to do so in a reliable and valid fashion, bolster the confidence to which the investigator is entitled when interpreting the results of the study.

In selecting a measure, it is important for the investigator to examine the available literature to identify the extent to which the scores on the measures have met the pertinent criteria for reliability and validity in ways that approximate the use in the present study. Many resources are available in print and on the Web to help obtain this information (e.g., Buros Institute, 2011; www.unl.edu/buros; ericae.net/testcol.htm). Also, merely typing in the name of the measure or "validity" or "reliability of" and then the name of the measure on a search engine (e.g., Google Scholar) is an easy and excellent place to start.

Table 10.1: Commonly Referred to Types of Reliability and Validity

Type	Definition/Concept
Reliability	
Test–Retest Reliability	The stability of test scores over time; the correlation of scores from one administration of the test with scores on the same instrument after a particular time interval has elapsed.
Alternative-Form Reliability	The correlation between different forms of the same measure when the items of the two forms are considered to represent the same population of items.
Internal Consistency	The degree of consistency or homogeneity of the items within a scale. Different reliability measures are used toward this end, such as split-half reliability, Kuder-Richardson 20 Formula, and coefficient alpha.
Interrater (or interscorer) Reliability	The extent to which different assessors, raters, or observers agree on the scores they provide when assessing, coding, or classifying subjects' performance. Different measures are used to evaluate agreement, such as percent agreement, Pearson product-moment correlation, and kappa.
Validity	
Construct Validity	A broad concept that refers to the extent to which the measure reflects the construct (concept, domain) of interest. Other types of validity and other evidence that elaborate the correlates of the measure are relevant to construct validity. Construct validity both on the theory underlying the measure (i.e., how that construct relates to multiple phenomena) and then concretely on the relation of a measure to other measures and domains of functioning of which the theory underlying the measure may be apart.
Content Validity	Evidence that the content of the items reflects the construct or domain of interest. The relation of the items to the concept underlying the measure. This is often evaluated by having experts in the area of the construct make judgments about the measure being relevant. This can also be evaluated statistically (e.g., factor analyses) to see how items go together and reflect content areas one would expect.
Concurrent Validity	The correlation of a measure with performance on another measure or criterion at the same point in time.
Predictive Validity	The correlation of a measure at one point in time with performance on another measure or criterion at some point in the future.
Criterion Validity	Correlation of a measure with some other criterion. This can encompass concurrent or predictive validity. In addition, the notion is occasionally used in relation to a specific and often dichotomous criterion when performance on the measure is evaluated in relation to disorders (e.g., depressed vs. nondepressed patients) or status (e.g., prisoners vs. nonprisoners). Does the measure correlate with the grouping variable and distinguish groups? Correlation (point-biserial), maximum likelihood estimates, and discriminant analyses are some of the ways to test for criterion validity.
Incremental Validity	This refers to whether a new measure or measure of a new construct adds to an existing measure or set of measures with regard to some outcome. That outcome might be in the present or future. Incremental validity is evident if the new measure adds significantly (statistically) and can be evaluated in multiple regression and discriminant analyses.
Face Validity	This refers to the extent to which a measure appears to assess the construct of interest. Not regarded as a formal type of validation or part of the psychometric development or evaluation of a measure. To say that a measure has face validity usually means no "real" (other types of validity are available). The value of face validity may stem from the likelihood that the measure is seen as a reasonable reflection of the domain of interest by persons who administer or who complete the test, but that alone does not attest to construct validity or any other type of validity.
Convergent Validity	The extent to which two measures that assess similar or related constructs correlate with each other. Convergent validity is supported if the measure correlates with other measures with which it is expected to correlate. The correlation between the measures is expected based on the overlap or relation of the constructs. Concurrent validity that takes on special meaning in relation to discriminant validity. Together these types of validity can help separate whether the correlations between measures are due to the relations of the constructs that are assessed or common method factors (all self-report measures) and their various combination. A multitrait-multimethod matrix (discussed later in the chapter) and various data analytic techniques are used.
Discriminant Validity	The correlation between measures that are expected *not* to relate to each other or to assess dissimilar and unrelated constructs. The validity of a given measure is suggested if the measure shows little or no correlation with measures with which they are not expected to correlate. The absence of correlation is expected based on the separate and conceptually distinct constructs.

NOTE: The types of reliability and validity presented here refer to commonly used terms in test construction and validation. See For Further Reading for resources that elaborate various types of reliabilities and validity.

In addition to selecting measures with reliability and validity data in their behalf, it is important to report on these characteristics when preparing a written report and it is important to look for information about reliability and validity of the measures when reading a report. A review of articles (>950) across seven health and behavioral journals revealed that authors very frequently fail to report on the reliability or validity of the measures they have used in the study (Barry, Chaney, Piazza-Gardner, & Chavarria, 2014). This is not merely a reporting issue but can mean there is no clear assurance that the constructs the authors were discussing in fact were suitably assessed by the measures.

10.1.6: Sensitivity of the Measure

The measure ought to be sensitive enough to reflect the type and magnitude of change or group differences that the investigator is expecting.

Measurement sensitivity refers to the capacity of a measure to reflect systematic variation, change, or differences in response to an experimental manipulation, intervention, or different group composition (e.g., as in a case-control study).

For example, if a study compared the effects of mindfulness training versus no training to reduce anxiety among persons visiting a dentist, a relatively large difference (effect size) might be expected between these two conditions. One might expect performance on the measure to be able to reflect that difference. If two versions of mindfulness were compared, the difference on the measure, even if there were one, might be more difficult to detect. Whether an effect or difference is obtained is in part a function of whether the measure can reflect differences and change, but of course also a function of what is being studied and compared.

Whether and how sensitive a dependent measure is to change or to group differences is difficult to specify in advance of a study because it depends on other details of the study (e.g., what the manipulation is, how strong or large the differences expected between groups). A few general desirable characteristics of the dependent measure can be identified:

1. The dependent measure should permit a relatively large range of responses so that varying increments and decrements in performance can be detected. If a scale has a narrow range (scores can only span from 0 to 10), the ability of the measure to delineate different groups or conditions may be a problem. Alternatively perhaps there are 10 items and each of these is evaluated by the subject on a scale from 1 to 7; the total across many items (e.g., 10 items) could then have a potential maximum score of 70. There might be categorical, yes-no questions that are not easily placed on a continuum (e.g., have you ever been pregnant, do you own a dog, and don't you just love methodology?). Here, the number of yeses across several items might be summed, on the assumption they represent the same construct of interest. General rule: we want our measures to have a healthy range from some lower score to some higher score so that groups, conditions, and people can be separated or differentiated by the measure. Single-item measures (that sometimes are used to assess a construct) and very short forms (often used to save time) of larger measures have a number of risks and problems (mentioned later) but one is that they do not provide a large range to discriminate among subjects exposed to different conditions or who show different levels of the construct of interest.
2. If participants score at the extremes of the distribution at pretest, this, of course, will only allow the investigator to detect varying degrees of change in the opposite direction at postassessment. If it is necessary to be able to detect change in only one direction, as might be the case in studies designed to compare two treatments both known to be effective, then the measure need not allow for bi-directional changes. In such a treatment study, individuals may be screened because they are extreme (e.g., high levels of posttraumatic stress disorder symptoms) and we expect and hope they get better from some effective intervention (but also probably through statistical regression). Yet, as a general rule, allow for bi-directional changes if possible or of possible relevance to the hypotheses. Even in a study screening for extreme scores, the experimental manipulation may have an opposite of the intended effects, at least for some of the participants. Assessing and evaluating these changes can be very important. In general, there should be some assurance in advance of the experimental manipulation that ceiling or floor effects will not be a limitation that could interfere with detecting differences among various experimental and control conditions. These effects restrict the spread of scores and could make the measure insensitive to real differences that exist.

Psychometric data for the measure and the possibility of a wide range for scores to vary are important, but it is also useful for the investigator to ponder the items a bit.

Often scales are used without really looking at the items carefully to see if it is reasonable to expect scores on the items to reflect change for a given group or differences between groups. Also, scrutiny of the items may lead to hypotheses about some portions of the scale (e.g., subscales, factors) that might be more sensitive to group differences than others and that may provide a more direct or specific test of the hypotheses. As the investigator ponders the contents of a scale, he or she may begin to think of alternative or additional measures to better test or elaborate the construct.

Overall, the sensitivity of a measure in an investigation should be assured the best one can prior to conducting the study. If a body of literature already shows the sensitivity of the measure to the manipulation or intervention or for group comparisons of interest, then preliminary work on this issue can be avoided. Many studies are conducted that closely build on or redress some ambiguity of prior research, and the evidence from prior research may be quite useful and relevant. If such evidence from prior and closely related research is not available, preliminary work before the full investigation might evaluate whether different manipulations reflect change on the measure. A small pilot study (e.g., 10–20 cases, 5–10 in each of two groups) can provide preliminary information about whether the measure could yield group differences (because all or most scores are not at the ceiling or floor of the scores). It is important to know whether the measure could reflect the predicted relation between independent and dependent variables. If no relation were demonstrated between the

independent and dependent variables at the end of the investigation, it would be reassuring to know that the reason for this was not the insensitivity of the dependent measure. An alternative to pilot work is to include the measure with several others on an exploratory basis and explicitly acknowledge in the investigation that one purpose is to explore the relation of a new measure with those already available in the literature. This latter alternative is a full-scale investigation rather than just pilot work.

10.1.7: Diversity and Multicultural Relevance of the Measure

There is another type of "sensitivity" that is required of measurement beyond the type just discussed. This has to do with whether or the extent to which measures used in a study are appropriate for, reliable, and valid for different groups within the study. Consider briefly background for this generally neglected consideration.

The population within the United States has changed and is changing markedly to reflect increased cultural pluralism and diversity. Currently, minority groups comprise 37% of the U.S. population, and this is projected to increase to comprise 57% by 2060 (~50 years) (United States Census Bureau, 2012). Currently, Hispanic American and Asian American groups are expected to increase from 17% and 5%, respectively, now to 31% and 8% in 2060. African Americans are projected to comprise roughly the same proportion of the population (14% currently and 15% in 2060). Non-Hispanic White Americans are projected to decline from the current 63% of the U.S. population to 43% by 2060. And still small in percentage but the fastest growing category is multiracial with African American-European Caucasian and Asian-European Caucasian being the two fastest growing groups within this category.

> The growing number of ethnic "minorities" is a critical point of departure for all human-related sciences. Certainly, we want psychological science to be relevant to the diversity of our culture and of the world cultures too.

Apart from the numbers, we know more now about the critical importance of culture. Culture and ethnic identity can be reflected in fundamental psychological processes (e.g., memory, perception, decision making). Also, central topics within clinical psychology such as rates and symptom patterns of psychiatric disorders, risk and protective factors, seeking of and response to psychological treatment, and merely to mention a few topics are influenced, sometimes greatly, by culture and ethnicity (e.g., Paniagua & Yamada, 2013). The centrality of culture and ethnicity has been recognized nationally as reflected in U.S. Surgeon General's Report, which noted that culture identities "affect all aspects of mental health and illness, including the types of stresses they confront, whether they seek help, what types of help they seek, what symptoms and concerns they bring to clinical attention, and what types of coping styles and social supports they possess. Likewise, the cultures of clinicians and service systems influence the nature of mental health services" (e.g., Satcher, 2001, v).

Increasing attention has been accorded culture and ethnic diversity in clinical domains. Prominent among the acknowledgment in the mid-1990s to evaluate the role of culture in the context of psychiatric diagnosis was a Cultural Formulation Model (see Lewis-Fernández & Díaz, 2002; Mezzich, 1995). The Cultural Formulation Model was devised to recognize, consider, and assess five components:

- Assessing cultural identity
- Cultural explanations of the illness
- Cultural factors related to the psychosocial environment and levels of functioning
- Cultural elements of the clinician–patient relationship
- Overall impact of culture on diagnosis and care

The focus on diagnosis in cultural context is critical in its own right. As one illustration, diagnosis of serious psychopathology (e.g., psychosis) is more likely when cultural factors are not taken into account (Adeponle, Thombs, Groleau, Jarvis, & Kirmayer, 2012). Re-diagnoses recognizing cultural issues and the context of symptom presentation change the diagnoses individuals receive. The overall point is recognizing that culture is not merely a little moderator but can affect such weighty topics as prevalence of disorders and response to treatment.

10.1.8: Core Features of Ethnicity, Culture, and Diversity

Ethnicity, culture, and diversity more generally are core features of what we study. I have highlighted three points:

1. Changing demographics of cultural and ethnic groups in the United States
2. Role of culture and ethnicity in psychological processes
3. Acknowledgment in many areas of clinical research including but well beyond psychiatric diagnosis

Traditionally, research on ethnicity and diversity within the United States and cross-cultural research as part of international studies have served as two areas not well connected. Each area begins with the view that culture can moderate many findings and that understanding differences and similarities of different groups and subgroups is a point of departure rather than an afterthought. In light of the importance of culture, it is essential to draw some of the implications for assessment.

First, culture and ethnicity include critical components that can influence the inferences we draw from measures.

A given measure (e.g., Beck Depression Inventory, Positive and Negative Affect Scale) does not automatically provide the same information across cultural groups. Interpretation of the items, threshold for responding to a given item or symptom, and so on are likely to vary. We will discuss various types of reliability and validity later but it is important to note now that these characteristics are not properties of a scale or measure. Rather, they are the properties of scores obtained in a particular context.

> Critical to that context is cultural identity of the sample. It cannot be safely assumed, without supportive data within a study, that a given measure is equivalent for different cultures.

When one is including diverse samples in research, it is valuable to bring to bear data within the study that supports how the measures operate or behave (reliabilities, validities) for different groups. We already know that cultural identity can serve as a moderator and those points to evaluating the hypotheses by cultural identity. The assessment point is slightly different. Is a given measure assessing the same construct and with the same psychometric properties for the different cultural subgroups within the sample? Providing data on that within the study would be an excellent addition and arguable someday might even be required.

Second, much further work is needed on establishing the construct validity of measures for diverse ethnic and cultural groups.

Ensuring a measure is appropriate to different groups is not merely a matter of translating measures (and back translating to ensure the content is addressed). The language alone does not ensure that the measure is equivalent across cultural groups (see Leong & Kalibatseva, 2013). I mention this here because developing and evaluating measures as a function of culture and establishing similarities and differences in meaning and responses and across the full age spectrum are understudied areas of research.

In relation to a given study, I mentioned presentation of data on reliability and validity for the sample, especially if the sample departs from those used in prior research. The reason is that reliability and validity cannot be assumed in any new application. The recommendation is now expanded based on cultural considerations. If there are subsamples within a given study, report the assessment data in preliminary analyses to convey that reliabilities and validities (available to report within the study) operate in a similar way among subgroups.

10.1.9: General Comments

Selecting measures for a study is often relegated to looking at the existing literature and seeing what other investigators have used. This is kind of drive through restaurant shopping for measures to get the meal and measurement selection over with. I noted this merely to insert pause in selecting the measure for a given study.

Measures that have been used frequently and appear to show the effects of interventions or group comparisons by other investigators continue to be used frequently as new investigations are designed:

- On the one hand, using a common or consistent set of measures drawn from the literature has the advantage of permitting comparison of results across studies. One can tell whether subjects were similar (e.g., in degree of depression) and whether the independent variable (e.g., mood induction) affects the dependent variable in roughly the same way (e.g., direction, magnitude). Common assessment methods across studies greatly facilitate such comparisons.
- On the other hand, much research is conducted in a tradition of weak or narrow assessment with little innovation to push or elaborate the limits of a construct. Precedence (used in a study that has been published) is a *de facto* criterion for measurement selection, but not one of the stronger criteria.

> As a quick guideline, ask a fellow researcher (or yourself), why are you using *that* measure?

If the answer begins with a comment that others have used the measure, this conveys the potential problem. There are important considerations in selecting measures, and prior use of the measure by someone else may or may not be one of them. But it should not be the first reason unless one is trying to replicate the prior study and may not even be a good reason without reassurances that original investigator traversed the thought processes highlighted here. The reasons ought to be based on:

- Construct validity
- Psychometric characteristics of performance on the measure
- Sensitivity of the measure
- Cultural considerations as they pertain to the sample included in the study

> At the end of the study, suitability of the measure may emerge as a basis for criticizing the results. Indeed, when hearing the results of a study, it is almost always meaningful, cogent, and important to ask, "But how was *x* [e.g., construct or dependent variable] measured?"

(It is important to name the dependent variable—my experience is that people look quizzical if you actually say "*x*".) The reason is that there may be little generality of the findings from one measure of a construct to another measure of the same construct, as I illustrate later. Also, it may be that the measure was great (by some psychometric criterion), but it is arguable whether it assesses the construct of interest.

10.2: Using Available or Devising New Measures

10.2 Examine the three avenues of choosing the appropriate measure in research

In most cases, the investigator will use available measures and report psychometric characteristics reported for samples used in previous research. When measures of the construct of interest are simply not available, however, the investigator may make the decision to develop a new measure to address the questions that guide the study.

10.2.1: Using a Standardized Measure

Many measures are available in an area of research, and there is usually tacit agreement that certain types of measures, modalities of assessment, and specific instruments are important or central. For example, in studying adult depression, an investigator is likely to include a self-report measure (Beck Depression Inventory) and clinician rating scale (Hamilton Rating Scale for Depression). These modalities and these specific instruments have enjoyed widespread use, a feature that does not necessarily mean that the measures are flawless or free from ambiguity. These scales are considered to be the most well-researched within this area, and performance on the scales (e.g., scores that relate to the degree of depressive symptoms and correlates among these different levels of symptoms) is quite meaningful among investigators. The frequent use of the measures has fostered continued use, and researchers embarking on a new study (e.g., evaluating treatment for depression) usually include one or both of these in the broader assessment battery.

Another reason for using standardized measures, of course, is the amount of work that may have gone into the measures by other researchers. That work facilitates interpretation of the measure. For example, to assess intellectual functioning or psychopathology among adults, one might rely on the Wechsler Intelligence Tests (different tests from preschool through adulthood) and the Minnesota Multiphasic Personality Inventory (MMPI-2), respectively. Thousands and thousands of studies of these measures with diverse samples and diverse cultures facilitate their interpretation. Also, use of such well-studied measures lends credence that a new study assessed the construct of interest. Similarly, fMRI has been used as a neuroimaging technique for some time (beginning in the early 1990s), and much is known about its use. There may be controversy about what can be concluded in any given study (e.g., what can and cannot be concluded from brain activation) (e.g., Bandettini, 2012; Lee et al., 2010). Yet, there are now fairly widespread accepted and commonly used methods of fMRI and software, scoring, and data-evaluation techniques to interpret and display the findings. In all of these circumstances, the investigator need not worry about defending the measure or providing support for the construct validity in light of the prior work that has been completed.

> Yet, there can be a trade-off. Does the standardized measure assess the precise construct or aspect of the construct of interest?

If yes, that is wonderful. If no, this means one might have a wonderful measure of the wrong construct.

The prior comments argue for selecting a well-researched measure when possible and to be sure that the construct of key interest is measured by that. That is all well and good, but an overarching tenet of methodology and science is to be skeptical and to question (but try to be nice about it). So even if a measure is standard and well-researched, that does not free it from your skeptical evaluation. The reason is the potential weakness that comes from being a standard measure.

> Standard measures take on their own life in the sense that once they are used a few times, there is a snowball effect in the accumulation and accretion of other additional studies.

Soon the measure is used automatically in an assessment battery without scrutiny. Occasionally researchers come along and scrutinize the data (e.g., psychometric properties) for the measure and convey the slightly embarrassing news that the measure is not all that great.

Using the Hamilton's Rating Scale for Depression is an excellent example because of its common and widespread use in depression research, as I noted previously. Scrutiny of the reliability and validity data from 70 studies spanning over three decades revealed that the psychometric properties and individual items across many samples are not that great at all and key types of validity (e.g., convergent and discriminant) are lacking (Bagby, Ryder, Schuller, & Marshall, 2004). These authors properly ask the question whether this measure, considered as the "gold standard" for assessing depression, is really a "lead weight" and something we might abandon for better measures.

The key point is not about the Hamilton scale but rather about use of standardized measures. Sometimes the comfort they provide is just as weak as what a child gives as an excuse on the playground when caught doing something and replies, "Everyone else is doing it too." When that child grows up, one hopes she does not select measures for a study using that same rationale. There might be lesser used measures that are just as good or better and one might vary the measure or develop a measure better suited to one's hypotheses.

10.2.2: Varying the Use or Contents of an Existing Measure

A standardized measure of functioning, cognitive processes, personality, behavior, or some other domain may be available, although some facet of the investigator's interest may make that measure not quite appropriate. The measure may have been developed, established, and validated in a context different from that of the proposed study. For example, one might wish to assess a geriatric sample, but the measure of interest has been developed, evaluated, or standardized with young adults. Alternatively, the investigator may wish to assess a particular ethnic group whose language, culture, and experiences differ from those samples with whom the measure was developed. The reason for selecting the measure is that the method or content seems highly suitable for the investigator's purposes. Yet, the measure has not been used in this new way or validated in the new context.

As I mentioned and invariably important to bear in mind, reliability and validity are not characteristics embedded in a measure. Rather, psychometric properties are related to scores of the measure in a particular use (e.g., sample, context). It is useful to know that a particular measure has yielded adequate to good reliabilities and validities across many circumstances and that is one reason to consider use of that measure in closely related but new circumstances (e.g., slightly different application from prior studies). Yet, the new use cannot assume adequate reliability and validity. It becomes more difficult to persuade oneself as a researcher or as a reader of a research study that the measure was fine (reliable, valid) in this new use as that use departs from those conditions that have already been well-studied.

If one is applying a tried and true measure in a new use, it is very helpful to include within the study some effort to evaluate psychometric properties in this new use. The task is to provide evidence that scores on the measure behave in a way that parallels the more common and standard use of the measure. Evidence regarding reliability is very useful, but greater concerns are likely to be voiced in relation to validity of the measure in its new use.

> Evidence might include correlating scores on the measure in its new use with scores on other measures in the study or using the measure to delineate subgroups and showing that the findings resemble those obtained in studies when the original measure has been used as intended.

If one is preparing a manuscript (e.g., for publication or equivalent paper), before presenting the main findings, often it is useful to include in the Results section preliminary analyses that evaluate the measure in its new use with any psychometric (reliability, validity) data that could be brought to bear. It may be sufficient to show that the new use of the measure leads to predictable differences on the dependent measure, although this may vary as a function of the complexity of the predicted relations and the plausibility of alternative interpretations of the results on the measure. Yet it is even better to show that and some psychometric properties associated with the new use. In the general case, it is advisable within the study or as part of pilot work to provide additional evidence that the construct of interest is still measured in the new use of the measure and that the measure still enjoys adequate psychometric properties.

10.2.3: More Information on Varying the Use or Contents

Use of existing measures in novel ways is often preferable to creating entirely new measures because the available research on the existing measure (e.g., original factor structure, correlations with other measures, and psychometric characteristics from various samples) is still relevant for interpretation of the measure.

If an entirely new measure were created instead, none of this background information would be available. On the other hand, use of standardized measures in novel ways may be viewed and labeled by colleagues who review the research as inappropriate or beyond the intention of the founding fathers and mothers who devised the measure. There becomes a point at which applicability of the measure to new samples, populations, and circumstances is strained and the challenge is appropriate. For many colleagues, that point consists of any extension beyond the specific purposes for which the measure has been developed and standardized. Reasonable people differ on this point, but reasonable investigators (you and I of course) provide some validity data to ally the cogent concern that the novel use is inappropriate or difficult to interpret. The validity data are not merely intended to allay concerns of others; we want to be sure that more than anyone else that we are studying the phenomena of interest as intended.

Investigators often make slight variations in a standardized measure such as:

- Deleting a few items
- Rewording items
- Adding new items

The purpose is to make the measure better suited to the new population or application. For example, questions asking about suicide attempt or violent acts may be omitted in a study of a community sample because the base rates of these behaviors might be low and the items would be potentially upsetting and provocative in that context. Approval of the research (e.g., Institutional Review Board of a university) may even require deletion of items of a scale. The same measure in a clinic setting would include

the items given the goal of identifying the full range of symptoms and the expectation that such items may be required. Omission of one or two items is a minimal alteration of the scale, and the items usually can be interpreted as if the scale were the original, by making changes in subscale or total scores (e.g., by prorating missing items or imputing missing data for that item in another way). Yet, this is all a matter of opinion, which is why we provide data to show that the scale still behaves in the same way.

There are little data available on the extent to which investigators make minor alterations in measures and the impact of these changes on the findings. Yet, available evidence indicates that "standardized," well-used measures are not really as standard as we thought. An evaluation of research using the Hamilton Rating Scale for Depression found that there are at least 10 distinct versions of the scale in use based on variations in wording and the number of the items (Grundy, Lunnen, Lambert, Ashton, & Tovey, 1994). (And additional variations have been used since this study was completed [e.g., Bent-Hansen & Bech, 2011].) Moreover, each variation did not have suitable reliability or validity data in its behalf or the strength of data that characterized the original version of the scale. It is likely that many researchers have lost track of the original scale, because as Grundy and colleagues noted, citations to the scale in a given study often are mistaken, i.e., they refer to a different version from the one used in the study.

> In short, standardized tests are likely to be altered; it is important to provide data that the altered version is as meaningful and valid as the results from use of the original version.

As a more general rule, when one tinkers with the content or format of a measure, the requirements are similar. As a minimum, some evidence is needed within the study to show the measure continues to assess the construct of interest and behaves psychometrically in a defensible fashion. To the extent that the measure is altered and that the new use departs from the one for which the measure was standardized, stronger and more extensive validity data are likely to be demanded by the research community.

As an illustration, in the work of our research group, we have been interested in measuring hopelessness in children in part because of work on an inpatient service where many admissions were children with depression and/or suicidal attempt. Among the issues that make hopelessness interesting is the relation to depression and suicidal attempt and ideation in adults, a topic that continues to gather research (Hirsch, Visser, Chang, & Jeglic, 2012; Klonsky, Kotov, Bakst, Rabinowitz, & Bromet, 2012). Hopelessness, or negative expectations toward the future, has been reliably assessed in adults with a scale devised for that purpose (e.g., Beck, Weissman, Lester, & Trexler, 1974) and frequently used as the Beck Hopelessness Scale (e.g., Hirsch et al., 2012; Neufeld, O'Rourke, & Donnelly, 2010). In developing the scale for children, the items from the adult scale were altered to simplify the content and to be more relevant to children's lives. Clearly such changes are not minor modifications of a scale but lead to qualitative differences in focus and content. Hence it is not very reasonable to assume that the original validity evidence obtained with adults would apply to children. Initial studies were conducted to provide reliability and validity data. Internal consistency data and analyses of items paralleled the results obtained with the adults scale. In addition, the construct of hopelessness in children generated results similar to those obtained with adults. Initial studies of the Hopelessness Scale for Children found that hopelessness correlated positively with suicide ideation and attempt and depression and negatively with self-esteem (Kazdin, Rodgers, & Colbus, 1986; Kazdin, French, Unis, Esveldt-Dawson, & Sherick, 1983; Marciano & Kazdin, 1994). Such studies are promising insofar that they support the construct validity of the measure and are similar to findings with adults.

Even so one or a few studies are limited, perhaps especially so if they emanate from one research program. In the case of our research, the children were within a restricted age range 6–13 and were all inpatients from a psychiatric hospital. Also, a limited range of constructs and other measures were examined to evaluate validity of the scale. In short, the studies provide some, albeit very incomplete, evidence regarding the new scale and how it behaves. The task in developing a measure is not necessarily to complete the full set of validational steps. Once an investigator provides preliminary evidence and places the measure within the public domain, others may complete further studies that greatly extend research on construct validity and psychometric issues, as is the case for the measure of hopelessness in children (e.g., Fanaj, Poniku, Gashi, & Muja, 2012; Merry et al., 2012; Phillips, Randall, Peterson, Wilmoth, & Pickering, 2013).

10.2.4: Developing a New Measure

Sometimes measures of the construct of interest are simply not available. The investigator may wish to develop a new measure to address the questions that guide the study. Instrument development can serve as a program of research in itself and occupy a career. In most cases, investigators are not interested in developing or evaluating a measure with that in mind. Rather, the goal is to address a set of substantive questions and to conduct studies that measure the construct in a new way.

Developing a new measure is a weighty topic in its own right in light of advances in measurement theory and scale construction and is beyond the scope of this chapter (see Kaplan & Saccuzzo, 2013; Wasserman & Bracken, 2013).

In developing a new measure, some evidence is required, either in pilot work reported in the write-up of the study or as part of the study itself, which attests to the validity of the measure.

The steps extend beyond face validity, i.e., that the content of the items is reasonable or obvious. Various types of reliability and validity, as presented previously in Table 10.1, might be relevant. Particularly crucial would be evidence that supports the assertion that the measure assesses the construct of interest. Such evidence might be reflected in one or more of the following:

1. Differences between groups on the measure (e.g., older vs. younger, clinically referred vs. nonreferred cases) in ways that are consistent with the construct (criterion validity)
2. A pattern of correlations showing that the new measure behaves as predicted, i.e., evidence that the direction and magnitude of these correlations are consistent (e.g., low, moderate, high) with what would be predicted from the relation of the constructs encompassed by the new and more established measures (concurrent, predictive, or concurrent validity)
3. Evidence that the new measure is not highly correlated with standardized measure of some other, more established construct (e.g., intelligence, socioeconomic disadvantage, social desirability), which might suggest that the new construct is fairly well encompassed by or redundant with the other (more established) construct (and does not meet discriminant validity)
4. Evidence that over time, performance on the measure does or does not change depending on the nature of the construct (e.g., mood vs. character trait, test–retest reliability)

With the use of a new measure, evidence on one or more types of validity is a minimum required to argue that the construct of interest is encompassed by the measure. As noted in the discussion of altering a standardized measure, it is usually insufficient to add the measure to the study and to show that it reflects changes that are predicted. Within the study, separate and independent types of evidence are needed about the measure apart from or in addition to how the measure reflects change as a dependent measure. However, the persuasiveness of any particular demonstration on behalf of a new measure depends on a host of factors (e.g., complexity of any predictions and clarity of the findings).

As an example from our own work at a clinic I have mentioned, we have been interested in why families drop out of therapy prematurely, i.e., early and against advice of the therapist. Actually, I was not very interested in this, but the topic was forced on me in doing treatment outcome research with children refer for severe aggressive and antisocial behavior. Rates of attrition in child therapy are high in general (40–60%), but are particularly high among families of children with aggressive and antisocial behavior for reasons not yet clear. Some of the factors that predict dropping out are well studied (e.g., low socioeconomic status of the family, parent stress, single-parent families). Variables such as these are helpful in predicting who drops out but not very informative because they do not shed light on why someone drops out and hence what might be done to reduce dropping out.

We felt that for many families treatment itself raises barriers or obstacles that influence who drops out. We developed a measure, called the Barriers to Participation in Treatment Scale, based on our experiences with parents and obstacles they report (Kazdin, Holland, & Crowley, 1997; Kazdin, Holland, Crowley, & Breton, 1997). Meetings with therapists generated all we could think of from our cases (a few thousand families) of why they dropped out. We converted several of these reasons to specific items and piloted this to see how these items relate to each other. Finally, we selected 44 items that reflected stressor and obstacles that compete with treatment, treatment demands, perceived relevance of treatment, and relationship of the parent and therapist. We added 14 items to assess stressors unrelated to treatment (e.g., job stress, moving residences, and alcohol and drug problems).

The construct we wanted to measure (stressors associated with treatment) may be explained in part by stressors in the parents' lives that have nothing to do with treatment. The parent and therapist separately complete the scale; both versions are designed to capture parents' experience in coming to treatment.

The results of initial studies showed that scores on the measures predicted dropping out of treatment and other measures of participation in treatment (e.g., canceling appointments, not showing up), that scores on the measure were not explained by other more easily assessed variables that also contribute to dropping out (e.g., lower socioeconomic status, stress, and others), and that stressors associated with treatment are not explained by other stressors in the lives of the families.

What do we know from these initial studies?

Probably only that the measure is worth pursuing further. The results are consistent with the construct and provide preliminary support. All sorts of questions remain about the scale, content, and correlates and only a few of which have examined (e.g., Kazdin & Wassell, 2000; Kazdin & Whitley, 2006; Nock & Kazdin, 2005). Developing a new scale begins the path of validation completely anew, and initial studies are only a very first step. Many investigations and investigators are needed to extend the construct validity and applicability of the scale, refine its meaning, and clarify its utility (e.g., Smith, Linnemeyer, Scalise, & Hamilton, 2013; Williams, Domanico, Marques, Leblanc, & Turkheimer, 2012).

10.2.5: General Comments

The strength, specificity, and very likely the value or utility of the conclusions from a study depend on interpretation of what was measured and the meaning of performance on the measures. If extensive evidence is available for the construct validity of the measure, which is usually the case for standardized measures, the burden of interpretation is a reduced. The burden is never eliminated even here because psychological measures by their very nature raise manifold issues about construct validity, external validity, and potential response biases (e.g., social desirability was one already mentioned). Intelligence tests, for example, tend to be the most well-studied psychological instruments. At the same time, the tests are surrounded in controversy related to their interpretation and use, such as:

- What is really measured by the scales?
- Is this a special type and at that a narrow type of intelligence because it best predicts how well people do in school?
- How does this relate to other types of intelligence (e.g., problem solving) or other cognitive processes (e.g., decision making)?

As I have noted, if extensive evidence is not available for a measure or if the use of a well-studied measure is novel, it is valuable to include some information about the psychometric properties of the scale in the new use. Of course, sometimes one might develop a new measure and here of course much more extensive information is needed to suggest that the new measure is reliable and valid in some critical ways. Even though the goal of the study might be to test this or that hypothesis, it is useful to add to that a side light to provide data about some facets of reliability and validity.

10.3: Special Issues to Guide Measurement Selection

10.3 Report the need to be cognizant of related issues while choosing the applicable measures

There are several issues to be aware of and alert to when selecting measures. Perhaps the primary issue is what modality of assessment will be used, i.e., what types of assessment (e.g., questionnaires, psychobiological measures). Here I discuss issues that can address broader issues relevant to selection.

10.3.1: Awareness of Being Assessed: Measurement Reactivity

Measures most frequently used in research are presented to participants who are well aware that their performance is being assessed. Such measures are said to be *obtrusive to denote that participants are aware of the assessment procedures*. Obviously, participants know some facet of their personality or behavior is being assessed when they complete a self-report questionnaire or are placed into a somewhat contrived situation in which their behavior is observed.

> Awareness raises the prospect that performance on the measure is altered or influenced by this awareness. *If performance is altered by awareness of the measure, the assessment is said to be reactive.*

It is not necessarily the case that subjects' awareness (obtrusiveness) influences their performance (reactivity). Knowledge of the purposes of the measures and motivation of the subjects, and no doubt other influences (e.g., response sets), contribute to reactivity. A few problems can result from relying on measures when subjects are aware they are being assessed.

One problem that arises is that reactivity is a *method factor*, i.e., a characteristic of the measurement that may contribute to the results or scores on a measure. When two measures are administered, their correlation may be due in part because they were both obtrusive and reactive. Essentially, subjects may respond in a similar way across the two measures. Response set is a concept that captures one type of assessment bias that can emerge.

> *Response set or style in measurement refers to a systematic way of answering questions or responding to the measure that is separate from the construct of interest.*

The set or style is a systematic influence on how the individual answers the questions and can interfere with obtaining the true score on a measure. Table 10.2 summarizes four recognized response sets for easy reference. As noted there, the first one is an *acquiescence response set,* which is *a tendency for individuals to respond affirmatively (true or yes) to questionnaire items*. This does not mean that an individual high on this response set will answer all items in one way, but there is a tendency to agree that is systematic. What that means of course is that scores on a measure include one's standing on the construct (e.g., high in altruism) but also one's response set. This would occur in a simple situation where all the

Table 10.2: Response Sets that Can Influence Responding When Subjects Are Aware that They Are Being Assessed

Response Set	Defined
Acquiescence	A tendency for individuals to respond affirmatively (true or yes) to questionnaire items
Naysaying	Tendency for individuals to disagree and deny characteristics. This is the "other side" or opposite of acquiescence
Socially Desirable Responding	Tendency to respond to items in such a way as to place oneself in a positive (socially desirable) light
End Aversion Bias	A tendency to avoid extreme scores on an item (e.g., 1–7 scale) even if those extreme score accurately reflected the characteristic

items are coded in one direction so that a yes consistently means high in the characteristic (altruism). One can "fix" this so that agreeing for some items but disagreeing for others are signs of high altruism. That is, some items are "reverse scored" and hence worded in such a way that saying yes to most items does not lead to a systematic bias.

> A more well-investigated response set is socially desirable responding and is not so easily addressed. *A social desirability response set is where individuals tend to answer items in the direction of placing themselves in a positive (socially desirable) light*.

Here responses selected on a questionnaire or other measure where subjects are aware of assessment are in the direction of trying to provide a positive impression. Although we tend to think that only self-report questionnaires might be vulnerable to such biases, other types of measures (e.g., projective techniques and direct samples of behavior) have been known for some time to show such effects as well (Crowne & Marlowe, 1964). Here no matter how the items are worded (reverse scoring, requiring agreement or disagreement) the individuals tend to select alternatives that make them look good or make a socially positive impression. This is understandable of course. If the test score will influence being selected (e.g., for a job, for being connected to possible soul mates for a matching Website, for psychiatric hospitalization, for a team), one might be expected to hold back on admitting to socially checkered behaviors, characteristics, and tastes. Of course you might not want to risk mentioning any illegal behavior (e.g., for extra money you moonlight by selling drugs or that your streak of successful shoplifting without being caught has passed 100) or perfectly legal but low frequency behaviors (e.g., you used to be vegan but now pretty much you are a raw meat person; for years now your family Thanksgiving dinners are at drive-through restaurants). Socially desirable responding goes beyond psychological assessment. A concern in social media is that individuals may place something on their "page" that might be socially damning and could actually harm selection (e.g., admission to some program, receiving an award).

10.3.2: More Information on Awareness of Being Assessed

I mention response sets because they are a potential influence or bias in assessment when participants are aware that they are being assessed and that awareness can systematically influence their performance (see Podsakoff, MacKenzie, & Podsakoff, 2012). Systematic biases can operate even on measures that might seem relatively immune. For example, people often misestimate their height and weight when self-report is compared to actual measurement, and these differences vary as a function of age, sex, and culture (e.g., Burton, Brown, & Dobson, 2010; Spencer, Appleby, Davey, & Key, 2002). An early finding on the topic found that women tend to underestimate their weight much more than do men; men tend to overestimate their height much more than do women (Palta, Prineas, Berman, & Hannan, 1982). Perhaps this finding would fit in with impression management in light of cultural pressures on the different sexes.

In general, interpretation of psychological measures can be greatly enhanced by using multiple measures that vary in reactivity (e.g., one reactive, another not). For example, the construct may be operationalized by a self-report measure, but also by direct observation of performance out of the awareness of the participant or in a contrived laboratory situation where the purpose of the study and the assessment situation is ambiguous. If similar results are obtained across such measures, the investigator has greater assurance that conclusions are not restricted to some aspect of the assessment method or influenced by a particular response set.

10.3.3: Countering Limited Generality

The use of obtrusive and reactive measures may *limit generality* of research findings. The problem of reactivity of assessment can be elaborated by discussing external validity more directly. Because almost all psychological research with humans relies on subjects who know that their performance is being assessed, one can legitimately question whether the results would be evident if subjects did not know their performance was being assessed. We take for granted that how subjects respond to our questionnaires about stress, social support, and other key constructs really identify performance, perceptions, or feelings outside of our experiment.

> It is reasonable to assume that obtrusive measurement (e.g., questionnaires in the lab) is *correlated* with real-life (unobtrusive) indices of the constructs. Yet we have little idea of whether the correlation is very high.

The generalization question in relation to assessment is, "how does the subject respond when there is no special assessment situation (e.g., my study)?" Examining this question invariably improves the quality of the study.

Several solutions can minimize or even eliminate entirely the influence of subject awareness on performance. These solutions vary as a function of the specific method of assessment. With self-report questionnaires and rating scales, the instructions given to the participants often are designed to increase their candor and to decrease the influence of reactivity. **One tactic** is to tell the participants that their answers to the test items are anonymous and that their individual performance cannot be identified. Of course, in most investigations these claims are accurate, although the participants may not believe them. In other situations, instructions may be provided to minimize the

likelihood that participants will answer the items in a particular way. Subjects are more likely to respond candidly and less likely to place themselves in a socially desirable light if they believe they cannot be identified.

> Another strategy to minimize the influence of subject awareness on performance is to add *filler* or *buffer items* on a given measure. The filler items are provided to alter the appearance of the focus or to make the measure appear less provocative or intrusive. In the process, the true purpose of the measure, i.e., the construct of interest, is obscured.

For example, a self-report measure of various psychiatric symptoms, criminal activity, or sexual practices might be infused with items about interests, hobbies, and physical health. The participants are aware of the assessment procedures, but the filler items may obscure or diffuse the focus that would heighten reactive responding. The filler items may soften the impact of the measure, and the reactions that might otherwise be prompted. Of course, the success of such items to obscure or attenuate the emphasis is a matter of degree; adding a few buffer items (e.g., do you get colds a lot, have you ever collected coins or stamps as a hobby) to a newly developed Scale of Tendencies toward Extreme Terrorism may not help very much.

Another solution is to vary what participants are told about the task and how it should be performed. For example, the purpose of the test may be hidden or participants may be told that their test responses have no real bearing on their future and will not be used for or against them. Extremely bright or suspicious subjects recognize that statements like this reflect that in fact this information will be used for or against them. (This is sort of like a doctor saying to a child that, "this will not hurt!" One only learns through development such a statement often is a clear signal that something will be painful. In defense of our doctors and parents who say this will not hurt, anxiety and subjective experience of pain can be greater if one is expecting pain; turning off that expectancy is likely to reduce these [e.g., Ziv, Tomer, Defrin, & Hendler, 2010].)

> Alternatively, participants may be told to respond to the items very quickly. The purpose of "speed instructions" is to have subjects give little attention to what actually is measured and hence not deliberate about the content or purpose of the items.

These instructional ploys may or may not be plausible to the subjects, depending upon the circumstances of testing and the exact facets of personality or behavior that are assessed.

> The use of computers, mobile devices, and Web-based measurement in psychological assessment has implications for reducing the reactivity of assessment. Computers permit participants to answer questions directly by responding to items presented on a monitor or screen, often in the comfort of one's own home (or work place).

The questions are presented, and answers are recorded automatically without a human examiner. As mentioned previously, computerized assessment, when compared with the measure administered by an examiner, often yields more information about sensitive topics such as (e.g., alcohol consumption, sexual problems). In addition, respondents often report favorable attitudes toward computerized test administration. In short, although computerized assessment is obtrusive, it may be less reactive. Similarly, mobile devices assess functioning in everyday life. At random times of the day, an individual may be "beeped" to answer several questions about emotional states. The regular assessment on multiple occasions within a day and across days and assessment in everyday life may reduce biases in responding, a speculative comment yet to be tested.

When reactive procedures are used because of the unavailability of alternative assessment devices, one of the strategies that might be adopted is to encourage participants to respond honestly as possible. Although this may be naive when participants have a particular interest in their performance in light of some goal (e.g., job procurement, discharge from a hospital), the overall approach may be sound. In many cases, such as evaluation of progress in therapy, it is usually in the best interests of the client to respond as candidly and accurately as possible. In such cases, this message may be worth elaborating to the respondents to obtain samples of performance during assessment that are as representative of daily performance as the measures allow.

Assessment occasionally consists of direct observation of behavior over an extended period. With such measures, different solutions have been sought to decrease the influence of reactivity. For example, when behavior is directly observed in a naturalistic situation such as the home or at school, there may be a novelty effect and the early data may not represent daily performance.

> Usually the first few days are needed to individuals habituate to the observers. It is assumed that after a period of time, obtrusive assessment will become less reactive over time and exert little or no influence.

Whether performance under obtrusive and unobtrusive assessment conditions is similar requires empirical evaluation. Even under ideal conditions of administration, the fact that participants are aware that their behavior is to be assessed might affect generality of the results. Possibly the results of an experiment have little bearing on behavior outside of the reactive assessment procedures.

10.3.4: Use of Multiple Measures

As a general rule, more than one measure ought to be used in a given study to assess the (or each) construct of interest. It is rare that a single measure captures the construct completely

or well. There are important exceptions where one measure is viewed as *the* critical index of the construct of interest and there is relatively little or no ambiguity about the measure and the construct it reflects. For example, survival (i.e., not dying) is often used as a dependent measure in research on diseases and their treatment (e.g., heart disease and cancer). The measure (mortality) usually does not raise epistemological questions ("how do you know they were really dead?") or methodological challenges ("does 'not breathing' *really* get at the central features of the construct?" "What was the test–retest reliability of the measure?"). Of course, definitional questions arise when discussing life and death in the context of personal, social, and ethical issues (e.g., abortion, termination of life support systems) but not usually in the context of assessment for research purposes.

Multiple measures of a construct usually are advisable in research. Use of multiple measures may make the study more complex in many ways (e.g., more measures for the subject to complete, more data scoring and analyses, potential inconsistencies in the results).

Even so, the recommendation is based on two considerations:

1. Most constructs of interest (e.g., personality characteristic, clinical problem) are multifaceted; that is, they have several different components. No single measure is likely to capture these different components adequately. Consider the construct of depression. Some components of depression are based on self-report. Individuals *report* that they feel sad, worthless, and no longer are interested in activities that were previously pleasurable. In addition, there are *overt behavioral components*, such as reduced activity and social interaction and changes in eating (more or less eating). Similarly, *psychobiological components* include changes in sleep electroencephalogram activity. These different facets of depression may overlap, but they are not likely to be so highly related that one is redundant. Any evaluation of depression in, say, a test of treatment would be incomplete if change were merely demonstrated in one modality. Single measures might well be fine if the problem or focus is highly circumscribed (e.g., enuresis, isolated fears, and specific habit disorders), if the measure is one that the world views as rather definitive (e.g., death, DNA profile) or sufficient (e.g., heart rate, pulse), or the goal is to address a single facet of a problem (e.g., blood pressure as an outcome among hypertensive patients). However, in most research, multiple methods ought to be used whenever possible, at least for the core constructs of interest or that reflect the primary hypotheses.
2. Multiple measures of a construct are helpful to ensure that the results are not restricted to the construct as assessed by a particular method and measure. Performance on a given measure (e.g., score, level of the characteristic) is a function of both one's *standing on that characteristic* (e.g., level of self-esteem) and *the precise method in which assessment is conducted* (e.g., self-report questionnaire, one questionnaire vs. another). In other words, the measure itself can contribute to the findings and conclusions.

There is a natural tension in a given research project between representing a given construct well (by using multiple measures of that construct) and including multiple constructs with only one measure each in a study. It is not reasonable to require participants to complete onerously long assessment batteries. So multiple constructs in the study cannot each be represented by multiple measures. Also, invariably there are little pressures here and there from advisors, funding agencies, graduate students, or oneself to add a measure to get at one more construct.

The tension is between breadth of coverage (many different constructs) versus thoroughness in assessing a given construct.

The extremes could be represented by 5 measures of a single construct in a study or 10 measures of 10 different constructs. The former is fine but does not allow one to relate measure of the construct to other constructs; the latter is not likely to assess any construct particularly thoroughly.

In general, a compromise is useful to consider. Identify the main constructs of interest or the constructs that figure most prominently in the hypotheses. Let us call these the primary constructs because they guide the study. Here it would be useful to measure the construct(s) with more than one method of assessment. This will have all of the advantages mention before especially if the measures of the same construct include different methods of assessment (e.g., self-report, direct observation). Other constructs in the study may be secondary in that they are of interest and may be important for control purposes (e.g., as covariates) and if needed represent these with fewer or one measure each. This is obvious a compromise because representing a secondary construct with one is not ideal. Yet, this makes many studies feasible because of limits of what can be asked of the participants.

10.4: Brief Measures, Shortened Forms, and Use of Single-Item Measures

10.4 Describe the implications of the terms brief measures, shortened forms, and use of single-item measures

The term *assessment battery* refers to all of the measures that will be used in a given study.

Multiple measures in the battery can be extensive insofar as the study assesses multiple constructs (e.g., stress, social support, psychopathology, marital satisfaction, and emotion regulation), and some of the constructs may have two or more separate measures (e.g., self-report, other report, physiological response). While we are doing our arm-chair planning of what we want to assess, the assessment battery quickly burgeons to a long list of measures. Yet, the amount of time required by the participant to complete the measures can become prohibitively long (e.g., a few hours). Moreover, the participant may be required to complete the assessment battery on a few occasions at different points in the study (e.g., immediately before and after the experimental manipulation or at various intervals in a longitudinal study). Is there any possible relief for the poor participant? There are three assessment options. I will highlight each one briefly and then raise considerations and cautions.

10.4.1: Use of Brief Measures

This first option is important to mention to make a critical distinction for this discussion.

> A brief measure means what it says. The measure may have relatively few items and not take long to complete.

There is nothing special to note here because the requirements of a brief measure area all those discussed previously, i.e., establishing that the measure has the needed psychometric properties, in sensitive to change, and so on. Sometimes very brief measures can be quite useful after having demonstrated these characteristics.

For example, to identify symptoms of Generalized Anxiety Disorder, one scale was devised with 13 items and then dropped down to 7 items (Spitzer, Kroenke, Williams, & Lowe, 2006). The items were taken from the psychiatric diagnostic system in place at the time and the symptoms that defined the disorder (DSM, IV). Patients were asked how much a symptom bothered them during the past 2 weeks. Each symptom was scored on a 4-point scale from not at all, several days, more than half the days, and nearly every day. The validity of the 7-item version (called GAD-7) was supported in several ways (e.g., concurrent validity predicting lost days of work with higher levels of anxiety, patient status as evaluated by physicians, prediction of the full diagnosis and others) and was consistent with construct validity of the measure. This is a good example of a study with a brief measure that provides a careful evaluation of its validity.

There are issues that can emerge in brief measures, valid or not, and I shall take those up in discussion of considerations. However, as an initial point of departure, a standardized scale that is also brief raises no inherent problems. Such measures may be readily used because they have addressed the essential criteria in their development, because construct validity is defensible, and because of the added practical benefit of requiring less assessment time on the part of the participant than some longer version.

10.4.2: Use of Short or Shortened Forms

Brief measures can be considered as standardized scales that just happen to be brief. Short or shortened forms can be an entirely different matter. These are measures that are derived from and abbreviated versions of a longer standardized scale. Here too the investigator has as a goal in mind alleviating the burden of assessment and using a shortened version of a standardized scale.

As one example, consider the Symptom Checklist 90-Revised (SCL-90-R), a very widely used (internationally) self-report measure designed to assess psychiatric symptoms among adults (Derogatis & Unger, 2010). The 90 questions are presented in a true–false format and cover a wide range of symptoms domains (e.g., anxiety, psychoses, and depression). The symptom domains represent subscales, but research supports the total score as the more defensible measure because of the very high intercorrelations of the scales (rs > .9) and high internal consistency of the 90 items when considered as a single domain (by summing all of the items). The measure is used often in research but of course as only one of multiple measures. Many different shortened versions have been used.

For example, a recent report compared 3 shortened version of the scale (27, 18, and 14 items) with the full 90 item version among 2,727 patients with affective disorders (e.g., Prinz et al., 2013). Subjects completed the SCL-90 and from that full set of items, scores on the shortened versions were extracted. That is, patients did not complete four separate forms—they completed the full-length version, and smaller versions were made post-hoc by extracting subsets of items. The main findings were that psychometric properties (e.g., various forms of reliability and validity) were similar across the four versions of the measure. Also, all versions reflected changes in symptoms over time about equally well as patients completed the assessment on separate occasions. By and large the short form performed well insofar as showing results as obtained in the full 90-item version of the scale. There are many studies of SCL-90 and many short forms (e.g., in one study alone, 11 different short forms were evaluated [Müller, Postert, Beyer, Furniss, & Achtergarde, 2010]). In general, their psychometric properties suggest that brief versions were fine in terms of several indices of reliability (mostly internal consistency) and validity (mostly concurrent validity with other measures) and are useful as a brief screener for symptoms.

10.4.3: Single or a Few Items

Occasionally, investigators add a measure of some construct that is one or a few items. This is a case where the investigator believes it would be important to measure something but wants a very brief assessment. In the usual case, the items that are used are based on face validity only—an unacceptable indefensible criterion from the standpoint of methodology. The reason is that one does not know what any measure really assesses (with rare exception such as "death") without the proper validity data. So we may call the item, a 7-point scale to measure "empathy," for example, but without considerable other information, we have no idea what is measured. If one is tempted to use a few items that are home-made, the work behind the measure is pivotal.

Consider an example of a well-designed study focused on the antecedents of youth (ages 9–14) drinking alcohol (Fisher, Miles, Austin, Camargo Jr, & Colditz, 2007). Youth (>5,500 from different locales in the United States) were assessed and then followed for 2 years to see who would take up drinking. Among the key hypotheses was that family who ate meals together would have children who were less likely to take up alcohol use 2 years later. This was a prospective longitudinal study, so the time line (family eating meals together at time one) could be examined in relation to the outcome of interest (drinking 2 years later). Youth who ate meals with their families were much less likely to take up alcohol consumption. Yet, one always looks at the measures. How was the key construct measured? At time one, one item was used to determine if and the extent to which youth ate meals with their families. Specifically, youth were asked to answer the question: "How often do you sit down with other members of your family to eat dinner or supper?" And could answer: never, sometimes, most days, everyday. Those who answered never or sometimes were compared with those who answered most days and everyday.

The issue: A single item was used to define the construct. Although it is so tempting and appealing to be completely satisfied with face validity, that is not enough. Strictly, we have no evidence that the item measures family meal time (e.g., as opposed to social desirability as one parsimonious possibility). Another item to assess parent alcohol use was also included and that too might be explained by another construct that how frequently parents really used alcohol.

I hasten to add the use of one-item per se is not the critical point. As often as not studies use two to four items in the same vein, i.e., to mention a construct of special interest, to invent items to do that, but to assume validity and reliability without any evidence at all. Yet, assessment validation is not a luxury—we cannot rely on face validity about what an item or couple of items "really" measure. Perhaps one would argue the other way—wait, what evidence is there that the one or a few items do not measure what we say. What evidence could one use to refute in the above example that the one item measured "family eating together." Yet, this is not how science works. The onus is on us as investigators to show what a measure does assess. One cannot just say we are measuring a construct, invent a few reasonable sounding items, and ask the rest of the world to prove our measure is not valid.

The lesson from this: Use measures that go beyond face validity. One- or a few-item measures often have only that in their favor—not always. In any study, we convey to the reader (but even before that, to ourselves) that the measure we are using is a reasonable reflection of the construct we are studying. "Reasonable reflection" is a loose way of saying—is reliable and valid. A single-item or a few items might be reliable and valid, but face validity is not suitable support for either of those.

10.4.4: Considerations and Cautions

The main cautions apply primarily to shortened forms and single-or a few-item measures rather than brief forms. Again the distinction is that brief forms have as their requirements all of the usual in relation to reliability and validity and their selection ought to be based on that evidence. As a starting point, perhaps the examples (e.g., GAD 7, SCL-90 shortened forms) raise the broader issue of why do we not use brief or shortened forms for all of the lengthy measures to which we subject participants? There are methodological answers:

1. The purpose of the use of a given measure may make short forms especially helpful.

   If one wants a measure for a quick screening to identify who will be included or who will be subjected to a more intense assessment battery, short forms are particularly helpful.

   Perhaps the study only wants to include people who are experiencing mental health problems (or who are not). Use of something like an abbreviated SCL scale might have a cutoff score for the initial screening. The brief scale could be administered to many people with little inconvenience for this purpose. So this first answer pertains to the design of the study (prescreening included) and purposes of assessment.

2. There are cautions and reasons not to use short forms or shortened scales. A main reason is that the range of scores on a short version is restricted, obviously, by definition. So if one can go from 1 to 90 on a scale and 1 to 25 on a short form, that range could have implications for the study. In fact, people do not usually use the full range of a scale, so 1 to 90 probably is not accurate nor is 1 to 25 as the real range. A smaller range is likely for both long and brief forms than the numbers suggest. That means a short form is actually shorter (in range) than one might believe.

As a general rule, when we are trying to distinguish groups (e.g., group differences after an experimental manipulation or group differences after two or more groups are selected), we want the larger range of scores to help separate the groups. A restricted range makes group separation more difficult and the longer measure usually is preferred. Also, if we wish to correlate the scores of various measures, the longer scale is more likely to be beneficial because of the finer gradations (larger range) of scores that are possible. Similarly, if we want to show change (from one occasion to the next), the larger (longer) version is more likely to be able to do that because a greater range of scores means people can move on the scale in more nuanced way.

3. Related to the restricted range issue is that measures often are used to group or classify individuals. One wants high, medium, and low (e.g., sensation seeking) on some characteristic, and a measure will be used to delineate groups. Here too we want the longer form as a general rule to allow the full spread of scores and to lessen the likelihood that individuals with a score in the middle (on a shorter version) might well have been in the low or high range if more items and more nuanced scores (longer version) were available.

   Finally, shortened forms may not be appropriate if there are multiple subscales or characteristics in a measure.

   I used the SCL-90 example because research almost always uses the measure to provide an overall total symptom scores. Yet, other measures of constructs (e.g., personality, intelligence) often include subscales and separate characteristics or abilities. Here abbreviated versions may become less useful as options because items for the subscales are already relatively brief. Rather than abbreviating the measure across all items, perhaps some of the subscales could be dropped from the study because they are not relevant. That of course would abbreviate the assessment battery but retain the scales of interest.

4. A problem with short forms can be that for a given measure there may be many such forms and hence very little reliability and validity data are available for any particular form. We want to use versions of a scale that have been applied extensively and with scores showing reliabilities and validities across samples. Short forms are often evaluated but often not.

10.4.5: More Information Regarding Considerations and Cautions

Many of the above comments about restricted range can apply to brief forms, shortened forms, and single (or a few) item scales. Yet, the single- or few-item scales warrant special comment. The range on such measures is restricted so that concerns related to that are heightened. Yet, the most critical feature is the usual absence of validity about what the items measure. Outside of psychological assessment, one or two items are commonly used (e.g., answer a survey of whether you are going to purchase a new smartphone in the next 6 months, or whether you liked the experience in shopping at a Web site). Yes, these single-items have their problems, but we are talking about other issues. In science (psychological science), we are interested in constructs and their operational definitions (measures) and the extent to which those definitions reflect the constructs. We are describing phenomena, building theory, testing predictions, and so on. This is a different agenda from surveys and marketing. In conducting or reading about research psychometric properties of the scale and sensitivity reflecting change are not methodological niceties—they are essential and proper testing and support for our predictions depend on the adequacy of the measures.

What is the final word on using brief and shortened measures? There is none but critical factors to keep in mind as we make decisions in any research project or evaluating the report of someone else's research.

> Methodology (like life) often is a matter of trade-offs. The gain in brevity of a measure is being able to add it to an assessment battery when a longer version is not feasible.

Thus one has a measure of the construct and can speak to the literature that utilizes that construct and measure. The cost could be in sensitivity of the measure to show differences and change and to attenuate various statistics (e.g., correlation, effect size) because of the restricted range. In any given instance, the considerations need to be weighed (see Smith, Combs, & Pearson, 2012).

If one has to prioritize, perhaps identify the priorities of the constructs that are being evaluated in the study.

What are the main constructs to evaluate the hypotheses?

These constructs perhaps ought to be represented with more than one measure each and with as well-validated measures as available. Secondary or ancillary constructs might be better explored with brief measures. This is not a carte blanch for selecting measures with no validity data at all. Perhaps the study could provide such data. However, if the results do not "come" out with unvalidated shorted measures or the few items that are homemade, the first rival interpretation is that the measures were limited. Exert the most extreme caution in inventing a couple of items to represent a construct. That can be the worst of all worlds (no validity, restricted range, and no defensible statement to make about what was measured). When you read an article that notes something like "we used three items to measure (insert construct here)," you are entitled to roll your eyes (or someone else's eyes) if that sentence is not followed by some statement providing evidence beyond face validity.

10.5: Interrelations of Different Measures

10.5 Identify three explanations as to why the results obtained through multiple measures may vary

Although the use of multiple measures is advocated as a general strategy for evaluating a construct, this strategy has a price. The main issue stemming from use of multiple measures is that the results may be inconsistent across measures (see Meyer et al., 2001). Some dependent measures may reflect changes or differences in the predicted direction, and others may not. When results vary across measures designed to assess the same construct, this makes interpretation difficult or at least more complex.

As an example, the longitudinal study of physical abuse, sexual abuse, and neglect has shown that youths who experience abuse in childhood (<11 years old), compared with nonabused children, show greater violence and criminal behavior 20 years later (Widom & Shepard, 1996). Interestingly, the results varied by how the key predictors and outcomes were assessed. Early physical abuse in children was identified in two ways (official records, self-report in adulthood), and violence was also assessed in two ways (records of arrest and reports of violent activity). Early physical abuse, as measured by official records, predicted later arrest for violence in young adulthood. However, self-reported physical abuse did not predict later arrests. Interestingly, self-report measures of early abuse predicted self-report of violent activity. The results suggest that common method components (i.e., predictors and outcome measures that share the same methods) are related. Measures predicted less well or not at all when methods varied. The strength of the study is assessing predictors and outcomes using multiple measures and methods.

Consider another example, where newly wedded couples (N = 135) were asked to report on their partner and satisfaction with their marriage (McNulty, Olson, Meltzer, & Shaffer, 2013). There was a self-report measure that assessed their view. There was also an implicit attitude measure in which individuals responded to positive and negative words quickly presented in relation to photos of their partner or other people. The reaction time of associating positive or negative words with the photos was used to measure how positive the individual felt about the relationship or partner. Implicit measures require quick reactions and get at views people may have that are out of consciousness. In this study, there was a measure of marital satisfaction that was conscious and one that was implicit. Noteworthy is that the correlation of the two measures was 0 or written with more drama r = 0.00. As interesting, over the next 4 years, marital satisfaction was measured every 6 months. The results showed that the implicit measure predicted marital satisfaction over time whereas the self-report measure did not. One should not go too far with the conclusion. Only one self-report measure of marital satisfaction was included (and not a standardized measure of that construct) and that is a weakness before generalizing to all measures.

> Even so, the lesson for this discussion is unchanged, namely, it is wise to use multiple measures to better assess a construct and to be sure or to reveal whether a given finding is not restricted to one measure.

10.5.1: Three Reasons for Lack of Correspondence among Measures

The failure of multiple measures to agree in a study is a "problem" only because of some traditional assumptions about the nature of personality and human behavior and the manner in which independent variables operate. Actually, there are many reasons to expect multiple measures not to agree. Three explanations of the lack of correspondence among measures pertain to the contribution of method variance in assessing behavior, the multifaceted nature of behavior, and the magnitude of a client's standing on the characteristic. Consider these briefly because they can help guide interpretation of a study when measures do not reflect the same pattern or results:

1. The lack of correspondence among measures of the same construct has emerged as a special topic in the context of using multiple raters or informants. For example, a couple may rate many facets of their marriage; children, parents, and teachers may rate the children's behaviors; and a patient and clinician may rate the quality of the therapy experience or therapeutic change. When different raters evaluate the "same" domain, correspondence of the separate views is often poor. For example, when multiple informants are used (e.g., children, parents, and teachers) to evaluate child behavior, the correlations of the different informants all rating the same target (child behavior) correspond between .2 and .4 (e.g., Achenbach, Krukowski, Dumenci, & Ivanova, 2005; De Los Reyes, Thomas, Goodman, & Kundey, 2013). Again it is important to be aware of the sometimes limited correspondence among informants. This is not necessarily a problem. Measures from both or all informants can be valid. Different informants sample different behaviors, experience those behaviors differently, and see individuals in different contexts (e.g., school, home, and work). As a more general issue, measures of the same construct may not invariably go together. This can occur when raters are not the "method" but different methods are used. In many ways, the absence of high agreement among measures, when it does occur, should not be surprising. Different methods of assessment present different conditions to the subjects, and these yield, generate, and promote genuinely different responses.

2. Many constructs we measure have several facets (are multidimensional) and not all of the facets may be present or go together. For example, several measures are available to assess depression. However, depression is not a unidimensional characteristic or simply sadness. Many characteristics can be identified involving affect (e.g., sadness), cognition (e.g., beliefs that things are hopeless), behavior (e.g., diminished activity), and biological symptoms (e.g., changes in eating and sleep patterns). Given the multidimensional nature of personality, behavior, and clinical dysfunction, as illustrated by depression, the lack of correspondence between measures is to be expected. Different measures of the same construct might simply emphasize or give different weight to different components, and little correspondence is quite understandable.
3. The lack of correspondence of different measures of the same construct may relate to the magnitude or strength of the characteristic, trait, disposition, or clinical problem. It may well be that different measures designed to assess the same characteristic of personality or behavior will co-vary as a function of the client's standing on the characteristic. For example, in the case of anxiety, clients may be measured on self-report, overt behavior, and psychobiological measures. The different measures may or may not correspond, depending on the magnitude of the client's anxiety. Perhaps clients who are overwhelmed by anxiety would score at a very high level on each of the measures. At the other extreme might be individuals who show absolutely no anxiety and score the equivalent of zero or "none" on each measure. These extreme groups (very high vs. no anxiety) may show a consistent or a more consistent pattern across measures.

   More generally, be alert to the prospect that multiple measures of a construct may not go together. There are ways to address this matter including combining measures when possible if they relate moderately to each other.

Also, using special statistical analyses (e.g., latent variable analysis, factor analysis) can address multiple measures of the same construct and clarify their relation to each other and to other predictors. These are beyond the present scope of the chapter but there are readily available resources (e.g., Bartholomew, Knott, & Moustaki, 2011; Duncan, Duncan, & Strycker, 2006).

10.6: Construct and Method Variance

10.6 Examine convergent and discriminant validity

It is useful to look a bit more analytically at scores on a measure. A participant's score is determined in part by standing on the construct (e.g., how depressed the person may be) and the measure that is used to assess that. The contribution of the method of assessment to the score for a given dependent measure can be seen by looking at some of the characteristics of measurement devices in general.

When a new measure is developed, it is important to establish that the measure correlates with other measures of the same construct (assuming some other measures and criteria are available) and that the measure does not correlate with measures of seemingly unrelated constructs (Campbell & Fiske, 1959).

Convergent validity, mentioned earlier (Table 10.1), was used to denote that independent methods of assessing a given construct agree with each other, i.e., correlate positively and perhaps in the moderate to high range. For example, if two measures that supposedly assess empathy correlate highly, this would be evidence of convergent validity. The fact that the measures converge suggests that they assess the same construct.

Discriminant validity was introduced to denote that a newly proposed measure should be distinguished from measures of other constructs; that is, show little or no correlation. We ought to expect a measure not to be related to other measures of entirely different constructs. Evidence that is consistent with this expectation supports discriminant validity. For example, a newly proposed measure of empathy probably would not be expected to be highly correlated with measures of other constructs, such as intelligence, social desirability, or anxiety. Evidence that all of these correlations were not very highly related to empathy would support discriminant validity of our measure. Of course, if high correlations were found between the newly proposed measure and a more established measure of another construct, this would suggest that the measures really were assessing similar characteristics, no matter what the two assessment devices were called. The new measure would be suspect to the extent that the results can be explained by a better validated measure whose construct validity (for some other construct) has been established.

The way in which convergent and discriminant validity can be examined is to conduct a study designed to evaluate the interrelations among various measures. Whether one measure converges with other measures of the same construct, of course, can be assessed by administering two or more measures designed to assess the same construct and seeing whether they correlate highly. The way to see whether a measure diverges from others with which it should not correlate is achieved by including measures of different or unrelated constructs and seeing whether they do not correlate or correlate only to a very small degree.

To obtain this set of correlations requires administering a number of measures to the same individuals. Some of the measures would be designed to assess the same construct or personality dimensions; some would assess different constructs or dimensions.

The correlations obtained to determine convergent and discriminant validity cannot be viewed uncritically. There is a

nuance here that is important to know. It is quite possible that correlations between two measures will be influenced not only by the construct that is being assessed but also by the method of assessment. For example, if two paper-and-pencil measures of empathy were administered, they might correlate rather highly. Is this evidence of convergent validity? Actually, it may be that the high correlation is due to the similarity in the method of assessment (two self-report scales) rather than or in addition to the construct being assessed. In addition, it is possible that a high correlation will be obtained between measures because of a common source of bias or artifact. For example, participants may respond in a socially desirable fashion across both measures, and this will be misinterpreted as convergent validity for the construct that the investigator originally had in mind. For the moment, the reason for the correlation between measures that use the same method is not as important as realizing that such a correlation is likely. And one criticism of a given study (that you design or read) is that all the measures were self-report. If the study is about correlations, a limit is that some of the correlations that investigator believes are very interesting might be due in part to the fact that the same assessment method was used.

10.6.1: Using a Correlation Matrix

To evaluate the contribution to the assessment method, it is important to include with the assessment of multiple constructs or dimensions some measures that rely upon different modalities or methods.

A **multitrait-multimethod matrix** refers to the set of correlations obtained from administering several measures to the same participants when these measures include more than one trait (construct) and method (Campbell & Fiske, 1959; Hox & Balluerka, 2009). The purpose of the matrix (a set of several correlations) is to evaluate convergent and discriminant validity and to examine the extent to which the correlations between measures are due to the similarities in the way the responses are assessed (method variance) rather than in what constructs supposedly are measured (trait variance).[3]

Consider in a bit more detail a real example from my own work that focused on the assessment of children hospitalized on a psychiatric inpatient unit. Two main constructs were assessed (depression and aggression). Children and mothers completed a questionnaire and an interview designed to assess depression and then a parallel questionnaire and interview designed to assess aggression. Consider this a study with two constructs (depression, aggression) and two methods of assessment (raters). All measures focused on the depression and aggression of the child.

Figure 10.1 presents the correlation matrix in the form that permits examination of convergent and discriminant validity and the role of the rater (method variance) in the

Figure 10.1: Correlations of Children and their Mothers for Measures of Depression

Correlations of Children and Their Mothers for Measures of Depression (Children's Depression Inventory [CDI] and Bellevue Index of Depression [BID]) and Aggression (Hostility Guilt Inventory [HGI] and Interview for Aggression [IA])

		Children (N = 120)				Mothers (N = 120)			
		Depression		Aggression		Depression		Aggression	
		CDI	BID	HGI	IA	CDI	BID	HGI	IA
Children[a]									
Depression	CDI								
	BID	0.62[c]							
Aggression	HGI	0.38[c]	0.42[c]						
	IA	0.41[c]	0.49[c]	0.26[b]					
Mothers									
Depression	CDI	0.10	0.24[b]	0.14	0.29[b]				
	BID	0.16[a]	0.27[c]	0.10	0.20[a]	0.71[c]			
Aggression	HGI	0.13	0.17	0.24[a]	0.24[a]	0.46[c]	0.56[c]		
	IA	0.03	−0.04	0.10	0.23[a]	0.26[b]	0.54[c]	0.60[c]	

correlations. First, several numbers are in diagonals (and not enclosed within a triangle). These are correlations when the rater is different but the construct and measure are the same. These are used to support convergent validity. (Because these measures were administered at a given point in time, these also happen to be measures of concurrent validity.) Look at the circled numbers in the first column (vertical) in the table. For child and mother scores for the Children's Depression Inventory the correlation (r) is .10. This of course means that the children and mothers do not agree very much at all on the level of the children's depression. Disappointing but not odd; we have known for a while that parent–child agreement on child symptoms is very low.

The solid-line triangles include correlations of measures completed by the same rater (children or mothers). Within the triangles, it is useful to look at the correlation is with the same rater but different constructs. The dashed-line triangles include correlations of measures completed by different raters. Together, the different triangles show there is a rater effect, i.e., correlations tend to be higher when the rater is the same no matter what construct was measured. The diagonal rows of numbers between the dashed line triangles are validity correlations (completion of the same measures by different raters) ap< .05, bp< .01, cp< 001.

Of greater interest are the correlations enclosed in the triangles. The solid triangles include measures completed by the same rater (child or mother). In general, the correlations within a given solid triangle correlate relatively highly. This means that measures completed by the same rater or informant share an important source of variance even if the constructs they rate are quite different. Indeed, correlations by the same rater who rates different constructs (depression, aggression) tend to be higher than correlations of the same construct (depression) by different raters. For example, child ratings of depression and aggression on two interviews (BID, IA in the Table) were correlated at r = .49 (same rater and measures but different constructs). This is higher than ratings of depression between mother and child which were correlated at rs = .10 and .27 (same construct different raters).

The example permits separation of construct (trait variance) from rater (method variance). The study and many others like it show the strong contribution of rater variance. That is, measures of different traits within raters often are more likely to correspond than measures of the same traits between different raters. This finding has been consistent across different informants, including children, parents, teachers, peers, and hospital staff as sources of information (e.g., De Los Reyes et al., 2013). Stated more succinctly, there is a strong method (rater) component that often pervades the ratings, even though children can be reliably distinguished in terms of their depression and aggression.

Assessment in experiments can profit from knowledge that both substantive and methodological characteristics of assessment devices contribute to a subject's score. If information is desired about a construct, it is important to use more than a single measure. Any single measure includes unique components of assessment that can be attributed to methodological factors.

> Evidence for a particular hypothesis obtained on more than one measure increases the confidence that the construct of interest has been assessed.

The confidence is bolstered further to the extent that the methods of assessing the construct differ.

10.7: General Comments

10.7 Review the need to ensure that the selected measure assesses the construct of interest

The discussion of special issues related to selecting measures can be highlighted by noting several questions that can serve as guides in designing a study or evaluating one that has been completed.

> **First**, is there evidence that the measure one has selected actually assesses the construct of interest?
>
> The measure may seem reasonable (face validity) or someone may have used it before (argument by authority). These reasons are not enough for psychological (scientific) assessment. We need more and that is what the various forms of reliability and validity address.
>
> **Second**, what are possible factors that influence people's scores on the measures, and could these affect our conclusions?
>
> Three factors that come to mind are response sets or styles (social desirability), conditions of assessment (reactivity), and unique features associated with the assessment method (e.g., self-report). These can be surmounted by using more than one measure and having those measures vary in the assessment method (self-report and overt behavior; or neuroimaging and other-report) and if possible the conditions of assessment (e.g., unobtrusive measures).
>
> **Third**, could method factors limit the conclusions of a study?
>
> Yes, if multiple constructs are assessed with the same method, any correlations between the measures could be heavily influenced by common method factors. For example, much research now is conducted via the Web (e.g., MTurk). Participants are asked to fill out a set of self-report measures of different constructs (e.g., depression, relationship satisfaction, views of this or that social

issue). And now the investigator correlates all of these. This is not a strong study in part because the correlation includes the relation of the constructs to each other but also the common method factor—all self-report. The correlations may be inflated because of the shared method factor that pervades all of the measures.

Fourth, is there a better way to address many of the issues? Yes, for a key construct in a given study (e.g., aggression, arousal), try to use two or more measures that use different methods. We want to know whenever we can whether the results of our study are restricted to a single measure (narrow stimulus sampling as a threat to external and construct validity). Measuring the construct in more than one way adds strength to a study. If the results are consistent across the measures, that very much adds weight to the conclusions. If the results are not consistent across measures, that suggests intriguing hypotheses and prompts a little theory as to what happened. That too is an important contribution.

Summary and Conclusions: Selecting Measures for Research

Selection of measures for research is based on several considerations, including construct validity, psychometric properties, and sensitivity of the measures to reflect changes or differences. Also, it is important to consider the sample for which the measure will be used and whether psychometric properties apply to the use you intend. Culture and ethnicity were discussed. As measures are used with diverse groups, it cannot be assumed that the measures have the same meaning or properties. It is critically important to consider individual measures whether they are best suited to the goals of the project you are designing. Psychometric properties from prior studies as relevant to what you will be doing (e.g., test–retest reliability of repeated testing in your study) are a guide. This stands in contrast to a frequent way in which a given measure is—it was used before by someone else. There are better reasons to use a measure or to do almost anything else in life than, "someone else did it before."

Standard or currently available measures are usually used in a given study because they have considerable evidence in their behalf and because as investigators we wish to have our findings relevant to a broader literature in which the measures are frequently used. Occasionally, investigators alter standardized measures to apply them to populations or in contexts in which the measures have not been used or intended. The measure may be used as is or modified slightly by rewording or omitting items. Investigators may develop an entirely new measure because a standard measure is not available or alteration of an existing measure would on prima facie grounds render this of limited value. If a measure is used in a novel way, altered in any way, or if a new measure is developed, it is essential to include validity data within the study or as pilot work to that study to support construct validity.

Special issues were discussed that also guide selection of measures. Awareness of being assessed was discussed and may be a common method factor across all measures within a study and influence the findings. Socially desirable responding was one influence that may be prompted by being aware that one is being assessed.

Use of multiple measures rather than a single measure was recommended because:

- Constructs of interest (e.g., clinical problems, personality, social functioning) tend to be multifaceted, and there is no single measure that can be expected to address all of the components
- Performance may vary as a function of the assessment modalities and devices used
- Individual's standing on a particular dimension or construct may be partially influenced by the method of assessment

In advocating use of multiple measures, the assessment battery can be prohibitive for the participant. Hours (days) may be required to complete the assessment battery. Brief, shortened, and single-item measures were discussed. Many cautions were presented too because the primary criteria to keep in mind in measurement selection were those outlined at the beginning of the chapter (evidence for construct validity, psychometric properties, and measurement sensitivity). Shortened measures and the single- or few-item measures often do not have the requisite data to recommend their use or to allow their interpretation.

The interrelation of multiple measures of the same construct was also discussed. Different measures of the same construct do not always go together. This has been well-studied in the context of multiple informants who evaluate the "same" target (e.g., their marriage, child,

"life"). There are reasons to expect measures will not invariably go together as they sample different facets of a construct or different perspectives.

Finally, construct and method variance were discussed. The purpose was to underscore the fact that the method contributes or can contribute to the results. Also, a common method factor (e.g., all self-report scales) can inflate relations among measures. Convergent and discriminant validity and the multitrait-multimethod matrix help to reveal the contributions of construct and method variance and to provide an example of how they can be separated. Also, the fact that method of assessment can contribute to the results argues for using multiple measures that vary in the method of assessment.

This chapter has focused on the considerations underlying measurement selection. Among the common themes was considering the use of multiple measures to operationalize a construct and to use measures of different methods when possible. The next chapter focuses on the different types of measures from which to use and special issues that each type of measure may raise.

Critical Thinking Questions

1. Think of a situation in which some characteristic is measured and the measure, situation, or circumstances are likely to be reactive, i.e., change how a person responds from what it might otherwise be. How might you measure that same characteristic that would minimize reactivity?
2. You have developed a measure of "courage," which is designed to reflect how brave a person is in difficult situations. Think about validity in the following ways. What other characteristics would courage be related to (concurrent validity, convergent validity) if a study were done looking at different personal characteristics? What other characteristics would you expect to NOT relate to courage at all or very little (discriminant validity)? The answers to both of these questions contribute to the construct validity of your measure.
3. What could be misleading by using measures that only have face validity in their behalf?

Chapter 10 Quiz: Selecting Measures for Research

Chapter 11

Assessment: Types of Measures and Their Use

Learning Objectives

11.1 Recognize that the characteristics of a measure determines its selection as a research method

11.2 Analyze the properties of objective measures as used in clinical research

11.3 Express the use of global ratings in clinical research

11.4 Review the properties of projective measures in clinical research

11.5 Examine the utility of direct observations of behavior in clinical research

11.6 Express the properties, pros, and cons of psychological measures in clinical research

11.7 Scrutinize how computerized, technology-based, and web-based assessment has helped in clinical research

11.8 Describe unobtrusiveness measures as used in clinical research

11.9 Examine how a modality can be best suited for certain clinical research

We previously discussed several considerations that are used to guide evaluation and selection of measures, as well as the use of multiple measures to assess constructs. There are several considerations that are used to guide evaluation and selection of measures, as well as multiple measures to assess constructs. The reason was the inherent limitation of any measure capturing all facets of a given construct. Also, there is a "method factor," which means that the findings obtained by measuring a construct in a particular way (e.g., self-report) can be due to the content or construct of the measure as well as this method or way in which that construct was assessed. In most circumstances, we would like to know that the results are not restricted to one way of assessing the construct of interest.

There are many different types of measures that are candidates to include in a study. In this chapter, I highlight several types of measures that are commonly used. The focus is on these types or modalities of assessment rather than listing or covering individual assessment devices. The types are described and illustrated. Key considerations that relate to using a particular type of measure are also discussed.

11.1: Type of Assessment

11.1 Recognize that the characteristics of a measure determines its selection as a research method

The diverse measures available in clinical research and the range of characteristics they assess would be difficult to enumerate, let alone elaborate here. Measures used in clinical psychology vary in many ways. Table 11.1 presents salient characteristics that vary among measures and that have implications for selecting measures. In a given study, it is useful to select more than one measure of the construct of interest and to select measures that vary in their methodological characteristics. The strength of any finding is bolstered by showing that it is not restricted to one measure and one method of assessment. There are exceptions of course where only one measure (e.g., perception of pain, activation of a brain network) is the entire point of the study, but in most instances the general recommendation remains.

Characteristics in Table 11.1 help to identify different types of measures that may be selected and major selection options. The type of measure or modality of assessment is a much broader way of distinguishing methods.

Table 11.1: Dimensions/Characteristics of Psychological Measures

Characteristic	Definition/Concept
Global/Specific	Measures vary in the extent to which they assess narrowly defined versus broad characteristics of functioning. Measures of overall feelings, stress, and quality of life are more toward the global side; measures of narrowly defined domains such as mood state or emotional regulation are more specific.
Publicly Observable Information/Private Event	Measures may examine characteristics or actions that can be observed by others (e.g., cigarette smoking, social interaction) or assess private experience (e.g., headaches, thoughts, urges, and obsessions).
Stable/Transient Characteristics	Measures may assess trait-like characteristics or long-standing aspects of functioning (e.g., personality, self-control) or short-lived or episodic characteristics (e.g., mood immediately after being subjected to a frustrating experience in an experiment).
Direct/Indirect	Direct measures are those whose purpose can be discerned by the client. Indirect measures are those that obscure from the client exactly what is being measured.
Breadth of Domains Sampled	Measures vary whether they assess a single characteristic (e.g., introversion, anxiety, risk-taking ability, or need for social approval), whereas others are aimed at revealing many different characteristics of personality or psychopathology (e.g., several personality traits or different types of symptoms within a single measure).
Format	Measures vary in the methods through which subjects can provide their replies such as true-false, multiple-choice, forced-choice, fill-in, and rating scale formats of self-report scales and inventories and extended narrative reports subsequently coded as in projective techniques.
Automated and Equipment-Based	Measures rely on special equipment or activities that capture key processes or activity usually outside of the awareness of the subjects. Measures that rely on eye-tracking or brain imaging are frequently used in clinical psychology to characterize clinical samples (e.g., children with autism spectrum disorder) or activities (e.g., emotional regulation).

11.1.1: Modalities of Assessment Used in Clinical Psychology

Modalities of assessment commonly used in clinical psychology and related areas of research such as counseling and the mental health professions are summarized in Table 11.2 for easy reference but discussed further in detail to convey their characteristics, strengths, and limitations. There are many modalities not all of which are sampled here (e.g., ability testing, neuropsychological tests).

11.2: Objective Measures

11.2 Analyze the properties of objective measures as used in clinical research

Objective has a special meaning in testing and refers to characteristics of the measure that explicitly specify the material that is presented (the items) and in the response formats that are required to respond to them. Questionnaires that measure ability, personality, and intelligence are prime examples to keep in mind of objective measures.

Table 11.2: Type of Measure or Modality of Assessment

Modalities	Definition/Concept
Objective Measures	Measures that explicitly specify the material that is presented (the items) and response formats that are required to answer them. Questionnaires that measure ability, personality, and intelligence are prime examples; they present the items or task and have only a fixed set of ways in which to respond to them (e.g., 1–7 point scale).
Global Ratings	Efforts to quantify impressions of somewhat general characteristics. They are referred to as "global" because they reflect overall impressions or summary statements of the construct of interest.
Projective Measures	Assessments that attempt to reveal underlying motives, processes, styles, themes, personality, and other psychological process. Typically, ambiguous stimuli are presented and that allows the clients to respond with their own answers without necessarily fixing or restricting the response alternatives.
Direct Observations of Behavior	Measures that assess behavior of interest by looking at what the client actually does. The overt behaviors may be sampled from how the client performs in everyday situations or in situations that are designed explicitly to reveal specific responses.
Psychobiological Measures	Assessment techniques designed to examine biological substrates and correlates of affect, cognition, and behavior and the links between biological processes and psychological constructs. The measures encompass many different types of functions (e.g., arousal of the autonomic system), systems (e.g., cardiovascular, gastrointestinal, neurological), and levels of analysis (e.g., microelectrode physiology that permits analysis of the response of individual neurons in the brain and brain imaging in response to tasks and activities in human and nonhuman animal research).
Computerized, Technology Based, and Web-Based Assessment	The use of computers and automated collection of information as well as scoring and evaluating that information. This includes many formats, including those from current technologies (e.g., smartphones, tablets) and Web-based.
Unobtrusiveness Measures	Unobtrusive measures are a type of assessment that are out of awareness of the person whose behavior or other characteristics are being assessed.

11.2.1: Characteristics

The questionnaires that are used in objective measures have fixed (unvarying) questions and clear ways of providing answers (e.g., on a scale from 1 to 7, yes-no), and scoring keys (add up all items or subsets of items in a particular way). The measures are not "objective" in the sense of providing information that is free from opinion or judgment, the more everyday use of the word "objective." Rather the term merely means that the measure is in a special form that allows for consistency in how it is presented, completed, and then scored.

> Self-report inventories, questionnaires, and scales illustrate objective measures and are worth highlighting because they are the most commonly used type of measures within clinical, counseling, and educational psychology.

And there are endless measures for children, adolescents, adults, and the elderly that sample a huge array of characteristics, such as self-control, empathy, happiness, anger, and altruism—and we go on with an endless list of:

- Attributes
- Traits
- Mood states
- Domains of experience

These measures require clients to report on aspects of their own personality, emotions, cognitions, or behavior. Typically such measures include multiple items that are designed to sample specific domains of functioning (e.g., depression, quality of life, social support) and often have extensive supportive data on the construct validity of these domains.

The widespread use of self-report measures can be traced to several factors:

1. Many states, feelings, and psychological problems are defined by what clients say or feel. People often feel helpless, self-critical, generally unhappy, or have a low self-esteem and self-report is a direct assessment of these feelings, thoughts, and perceptions.
2. Self-report measures permit assessment of several domains of functioning that are not readily available with other assessment techniques. The client is in a unique position to report upon his or her own thoughts, feelings, wishes, and dreams, and overt acts and can report on his or her states and behaviors across a wide range of different situations and hence can provide a comprehensive portrait of everyday performance.
3. The ease of administration has made such measures especially useful for purposes of screening. Screening refers to the initial assessment phase where the investigator must select a small sample of cases from a larger population.

   > Often a simple assessment device (e.g., self-report scale) is used as a means to divide the sample. Individuals who meet particular criterion levels on the self-report measure or questionnaire can be selected and studied more intensively through other techniques.

There are many different types of self-report measures—so many that it is difficult to consider them as part of a single category. For many self-report measures, extensive research exists. For example, one of the most widely investigated measures in clinical psychology is the Minnesota Multiphasic Personality Inventory, an objective self-report test, which has been the topic of more than 12,000 books and articles and has now spanned research for a period of several decades. The revised version (MMPI-2) includes 567 true-false items and multiple scales that assess different facets of personality and psychopathology (Butcher, Graham, Williams, & Ben-Porath, 1990; Nichols, 2011). The measure often is used in its entirety, but several of its subscales have been used and validated separately (e.g., to measure alcoholism, depression, anxiety). The overall scale has been and continues to be used with diverse populations (e.g., patients with psychiatric disorders, prisoners, athletes) and for multiple purposes (e.g., screening of prospective employees, treatment planning, evaluation of therapy outcome, and even graduate student admissions).

Apart from any single measure, an extraordinary large range of measures designed to assess an overwhelming number of:

- Characteristics
- Traits
- States
- Moods
- Feelings
- Impulses
- Strivings
- Trepidations

Self-report measures can assess diverse aspects of a given characteristic or multiple characteristics merely by having the client respond to many different items. The number of measures available and the number of attributes, experiences, and personality characteristics that can be assessed make self-report measures very convenient and widely used.

11.2.2: Issues and Considerations

There are two general categories of methodological issues to raise in relation to self-report questionnaires. First, responses to items can be greatly influenced by the

wording, format, and order of appearance of the items. Rarely have these influences been studied in relation to measures used in clinical research. There are exceptions. For example, interviews are used to obtain psychiatric diagnoses of individuals covering multiple symptoms so that a psychiatric diagnosis can be derived. One does not think about this very much because the order in which the sets of items that cover symptoms of a particular disorder are presented to the subject is a bit arbitrary. Consequently, the impact of this order is rarely studied. Yet, we have known for some time that when self- or other-report diagnostic measures are administered, the order in which the disorders are assessed influences significantly the number of symptoms the patients show, whether they meet diagnostic criteria for particular disorders, and how impaired they appear to be (e.g., Franke, 1999; Jensen, Watanabe, & Richters, 1999). The findings do not seem to have influenced the development or administration of diagnostic instruments.

As researchers we may insufficiently appreciate the extent to which self-report responses are vulnerable to minor changes in how the items are worded, the format of the question, and the context in which any particular item is embedded. Consider a few of the many findings to convey the point (see Schuldt, Konrath, & Schwarz, 2011; Schwarz & Hippler, 2004). The extent to which people:

- View themselves as successful in life varies as a function of whether they rate this on a scale from 1 to 10 or from –5 to +5 (people rate themselves as more successful with the –5 to +5 scale).
- View marital satisfaction as a contributor to their overall life satisfaction varies as a function of the order of presenting questions about each type of satisfaction (i.e., marriage and life). Say they experience a variety of physical symptoms depends on how the response alternatives are noted on a continuum; 62% say they experience symptoms when the scale goes from "twice a month or less to several times a day," whereas only 39% respond with this frequency when the scale goes from "never" to "more than twice a month."
- View climate as an issue or problem is influenced by whether the questions use the term "global warming" or "climate change."

These findings only sample many similar results that convey that responses can vary markedly with format changes but also word changes.

> Subtle changes in item wording and the placement of a particular item in the context of other items greatly alter the responses.

Placement of the items also can influence key psychometric properties of the measure. For example, items toward the end of a questionnaire correlate more highly with the total score (minus that item) than items at the beginning of the test, holding constant the specific items. Also, if similar items are grouped together (e.g., items that measure anxiety, stress) rather than interspersed, the items that go together are more highly intercorrelated than they are when they are interspersed (mixed) throughout the measure. Thus, the structure of the scale (factor analysis) is clearer when the items are grouped on the measure itself (Knowles & Condon, 2000). It appears that early items, serial position of various items, organization (grouping) of the items, and response formats help subjects discern meaning of what is and is not being measured.

Even when items are constant in how they are presented, the answer can vary systematically by subject characteristics beyond the specific characteristic that is being measured. Fundamental interpretative, memory, and perceptual processes are involved that shape the answers (Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 2004). For example, answers to questions can vary as a function of culture and ethnicity because of how items are interpreted and response styles that can vary by culture (e.g., Jang, Kwag, & Chiriboga, 2010; Morren, Gelissen, & Vermunt, 2012). Cultural and ethnic issues are critically important because differences (e.g., in happiness, depression, use of substances) could be due to differences in the characteristic that is assessed or in how different groups would approach and interpret the item, or some combination of these influences.

Research on format and related structure of objective measures underscores a more general methodological point. Several seemingly minor facets of the questions, their format, and order of presentation can have impact on the responses.[1] In general, it is good that our standardized measures are used in a consistent way across investigators so that the order of the items, subscales, and other domains (e.g., disorders) is constant. At the same time, it is instructive to note that the results and substantive conclusions we reach dictated in part by the structure of the measure.

11.2.3: More Information on Issues and Considerations

An additional feature to note in relation to objective measures is the possibility of bias and distortion on the part of the subjects.

> *Distortion refers to the alteration of participants' responses in some way in light of their own motives or self-interest.*

This is error because the distortion is not measuring the construct of interest. And the error is systematic (not random) in the sense that it leads to responses in a particular direction. At the extreme, participants can dissimulate

to such a degree that the answers they report are simply untrue. Occasionally, inventories have special scales (e.g., lie scales) to assess the extent to which the subject is not telling the truth, is being inconsistent, or is endorsing response alternatives that are extremely unlikely.

Blatant dissimulation aside, subjects are likely to alter slightly the image of themselves that they present and to interpret very loosely the meaning of the items so that they appear to place themselves in the best possible light. As mentioned previously, the tendency to do this is referred to as *social desirability* and has been shown to be extremely pervasive on self-report measures. Long ago we have learned that inventories designed to measure psychiatric symptoms and personality traits often correlate very highly with measures of social desirability (e.g., Edwards, 1957). Thus, individuals who complete self-report items are likely to endorse the socially condoned behaviors rather than the socially inappropriate behaviors (symptoms). The pervasiveness of social desirability as a response style has led investigators to posit a specific personality trait referred to as the *need for social approval* (Crowne & Marlowe, 1964). This means that social desirability is not merely a response set that influences placing oneself in a good light on a particular measure, but has broader implications.

> Individuals who are high in their need for social approval on a self-report measure behave in experimental situations in a way that maximizes approval from others. Thus, the bias on self-report inventories has behavioral correlates beyond the testing situation.

Socially desirable responding and need for social approval as an influence on self-report measures remain topics of research, especially in research on personality (Paunonen & LeBel, 2012; Ventimiglia & MacDonald, 2012). Interest has continued in part because socially desirable responding is of interest in its own right and because it relates to other key constructs of interest within clinical psychology (e.g., emotional regulation, self-esteem, anxiety, participation in treatment, addictive behavior) (e.g., Arndt, Hoglund, & Fujiwara, 2013; Uziel, 2010; Zemore, 2012). Hence, socially desirable responding is not merely considered as a bias or methodological nuisance of self-report measures. Also, findings often are intriguing. For example, priming individuals to think about God increases their socially desirable responding, a finding that is interpreted as support for a supernatural monitoring hypothesis, i.e., the belief that one is being monitored (Gervais & Norenzayan, 2012). Thus, one can experimentally manipulate and alter socially desirable responding. There are other response set as listed in Table 11.3 that can operate (acquiescence, naysaying, end-aversion bias), although social desirability probably has received the greatest attention.

Table 11.3: Response Set

Response Set	Description
Acquiescence	A tendency for individuals to respond affirmatively (true or yes) to questionnaire items
Naysaying	Tendency for individuals to disagree and deny characteristics. This is the "other side" or opposite of acquiescence
Socially Desirable Responding	Tendency to respond to items in such a way as to place oneself in a positive (socially desirable) light
End Aversion Bias	A tendency to avoid extreme scores on an item (e.g., 1–7 scale) even if those extreme score accurately reflected the characteristic

The response sets are not minor methodological annoyances but can change how we view differences or lack of differences among groups. For example, we know there are many cultural differences based on how different cultures perceive, codify, and categorize experience. No one doubts that cultural differences are genuine. However, it is interesting to note in passing that some of the differences can be due to different response biases. In some cultures, respondents prefer to select extreme categories among response alternatives, whereas respondents from other cultures actually avoid extreme categories. These differences in response sets lead to substantive differences in how the cultures appear in their answers (e.g., Morren, Gelissen, & Vermunt, 2013).

In intervention research, another source of bias that has been discussed, but not well studied, pertains to changes in severity of symptoms that have little to do with genuine improvements. Before psychotherapy, clients may exaggerate their complaints because these exaggerations may ensure that they receive treatment or increase the speed with which treatment is provided. After therapy, clients may respond to the same measures in a more socially desirable fashion in the sense that they provide the therapist and clinic with evidence of improvement, presumably the reward of providing treatment.

> The changes in self-report responses before and after the therapy due to exaggeration and underplaying of problems are referred to as the *hello-goodbye effect* (Hill et al., 2013; Streiner & Norman, 2008).

I mentioned this previously in discussing instrumentation (and response shift) as a threat to internal validity, i.e., changes in some facet of the measure from one occasion to the next. Of course, this effect is difficult to estimate because of the actual changes in treatment or because of influences such as statistical regression, i.e., improvements that may result simply from having extreme scores at the initial assessment.

The problems of distorting answers on self-report inventories (e.g., socially desirable responding) stem from the fact that the subjects are aware that they are being assessed and may act differently than they ordinarily

would respond without this awareness. Participants bring to bear their own motives and self-interest in responding. The extent to which distortion may occur is a function of many factors, including whether subjects can detect the purpose of the measure and whether their motives are consistent with those of the investigator.

> Having clients complete tests under conditions of anonymity, ensuring confidentiality, providing incentives for candor, and conveying to the client that his or her best interests are served by honest self-evaluation are designed expressly to evoke more honest answers.

Even so, it is naïve to assume that the investigator's motivations for obtaining candor will override the subject's motives for self-protection and self-enhancement or concern about consequences of specific responses.

The pervasive use of self-report inventories and questionnaires in part derives from their ease of use. Also, self-report inventories have been extensively validated and shown in many instances to relate to nonself-report (e.g., behavioral, neuroimaging) criteria. Even in cases where we might expect maximum bias or distortion, meaningful validation data are provided. For example, if we ask parents to complete a measure that assesses the likelihood that they physically abuse their children or ask adolescents to report the extent to which they engage in delinquent behavior (e.g., DiLillo et al., 2010; MacDonald, Morral, & Piquero, 2011), we would expect socially desirable responding and denial as a rule. Yet, quite reliable and valid data have been generated that relate specifically to other criteria (e.g., measures of behavior, archival records of crime). Thus, the use of self-report is not merely a matter of convenience in selecting measures. The primary concerns are the pervasive use of self-report measures and often sole reliance on self-report as a method of assessing the construct or domain of interest.

Self-report merely illustrates the category of objective measures and suggests that clients or participants are the primary or sole source of data. Actually, clinicians, spouses, parents, and teachers often complete objective measures to rate a particular client or subject. Among the many features of objective measures is their amenability to studies that attest to the many types of reliability and validity. Relations of items with each other, development of subscales, and correlations of scales with other indices of functioning, all have permitted careful validation of many objective tests.

11.3: Global Ratings

11.3 Express the use of global ratings in clinical research

Objective tests refer to questionnaires and scales as I have mentioned. In clinical psychology, there are other measures based on reports that are not usually included as objective measures. They do not include the usual set of items that measure a construct. Global ratings are highlighted here because of their use in clinical research.

11.3.1: Characteristics

Global ratings refer to efforts to quantify impressions of somewhat general characteristics.

They are referred to as "global" because they reflect overall impressions or summary statements of the construct of interest. Typically, ratings are made by the therapist or by significant others who are in contact with the clients. A major justification for use of these ratings is that select individuals other than the client may be in a position by virtue of expertise (e.g., therapist, staff member of a psychiatric hospital) or familiarity with the client (e.g., spouse, parent) to provide a well-based appraisal.

The judgments may vary in complexity in terms of precisely what is rated. Very often global ratings are made in such areas as overall adjustment, improvement in therapy, social adequacy, ability to handle stress, and similar broad concepts. These ratings usually are made by having raters complete one or a few items rated on a multiple-point continuum where the degree of the rated dimension can be assessed. For example, a typical item might be:

- To what extent has the client improved in therapy? (check one)

___	___	___	___	___	___	___
1	2	3	4	5	6	7
no improvement		moderate improvement			very large improvement	

or

- How much do the client's symptoms interfere with everyday functioning?

___	___	___	___	___	___	___
1	2	3	4	5	6	7
not at all			moderate			very much

The preceding samples not only illustrate a commonly used format for global ratings but also the generality of the dimension frequently rated. Usually ratings ask for an appraisal of a multifaceted or complex area of functioning.

Global ratings provide:

1. A very flexible assessment format and can include virtually any construct of interest (e.g., symptoms, overall functioning, and comfort in social situations). The flexibility also means that a general characteristic can be used to rate individuals who may differ greatly in their individual problems. By rating clients on a global dimension that encompasses diverse problems (e.g., degree of improvement, extent to which symptoms interfere with ordinary functioning), a similar measure can be used for persons whose characteristics at a more molecular level vary greatly.

2. A summary evaluation of a client's status. The problems clients experience may include many facets (e.g., all sorts of symptoms or disorders). It is important to determine with specificity how these different facets have changed, but also useful to have an overall statement that distills the effects of treatment into a relatively simple statement (e.g., are you better off now than you were when you came for treatment).

   Global ratings also provide a convenient format for soliciting judgments of experts, peers, or other informants.

Presumably, an expert in the nature of clinical dysfunction is uniquely skilled to evaluate the status of the client, the severity of the client's disorder, and the degree to which change, deterioration, or improvement has occurred. Similarly, individuals in everyday life who interact with the client (e.g., peers, spouses, employers) also are in a unique position to evaluate performance. In this context, global ratings of client change often have been incorporated into treatment evaluation.

For example, the Global Assessment of Functioning Scale, used to assess the overall functioning on a mental health-illness continuum, consists of a single item from 1 to 100 (American Psychiatric Association, 1994). Broad descriptive guides are provided at 10-point increments (e.g., 1–10, some danger of hurting self or others; 51–60, moderate symptoms or moderate difficulty in social, occupational, or school functioning; 91–100, superior functioning in a wide range of activities). Multiple constructs and domains are interspersed on the continuum (e.g., symptoms, interpersonal relations, work and school functioning). The rating scale is global in the sense that one summary item is designed to represent how one is doing in life. The scale is global in another sense. The measure is still in active use in studying clinical dysfunction throughout the world (e.g., Aas, 2011; Kirpinar & Oral, 2012).

In a larger battery where specific constructs are assessed, one may want to include a global rating scale. After all is said and done and after one has addressed changes or group differences on the main constructs of interests, we may want to know answers to such questions as:

- Do the clients feel better?
- Do they see life differently?
- Do they relate better to significant others?
- Do they experience their overall functioning and impairment as better?

These are global questions, but the global questions of life are not trivial. In using measures to address them, it is important to ensure that other measures are also included to better evaluate the critical constructs that the investigator may wish to talk about in explaining the findings. Also, it is useful to evaluate the global ratings within the study (e.g., to correlate them with other measures and to regress other variables onto them) to facilitate interpretation of what they measure and mean.

11.3.2: Issues and Considerations

One of the major problems with global ratings is evaluating precisely what they measure. The phrasing of global ratings *suggests* what the item is designed to measure (e.g., symptoms). However, there is no assurance that this in fact is what is actually measured (see Aas, 2010). Few or, more often, no concrete criteria are specified to the assessor who completes the ratings. By definition, the ratings are rather general, and all sorts of variables may enter into the rater's criteria for evaluating the client.

Because the criteria are not well specified, it is possible that the global ratings may show changes in client behavior over time independently of whether the client has changed, as reflected on some other, more specific measure. For example, therapists may view clients as improving over time simply because of changes in the criteria used in making their overall ratings of improvement.

> Thus, a client's greater ease, candor, or warmth within the therapy session may influence a therapist's rating of client improvement at the end of therapy whether or not clinical change in the problem area (e.g., obsessions or compulsive rituals) has occurred.

Changes in the measurement procedures or criteria over time were referred to previously as instrumentation, a threat to the internal validity. Instrumentation can account for changes over time as a function of assessment (procedures, definitions, or criteria) rather than change in client behavior. Global ratings are especially vulnerable to the instrumentation threat because the criteria that go into making ratings are general and varying definitions are fostered by the generality of the items or questions.

Another problem with global ratings, certainly related to the problem of what they measure, is their potential lack of sensitivity. Essentially, global ratings ask the general question for a given dimension, such as "how severe are the client's symptoms, how much improvement has there been, and how anxious is the client?" By posing general questions, the measures lose some of the sensitivity that could be obtained from assessing very specific characteristics of the relevant dimensions of interest. Global ratings greatly oversimplify the nature of functioning and therapeutic change. By utilizing a global measure, the richness of detail is lost.

The strengths and limitations of global ratings can be illustrated by study of psychotherapy that prompted years of debate in clinical psychology. Approximately 20 years ago, a survey was administered by *Consumer Reports* (1995) that asked adults to report on the extent to which they were satisfied with psychotherapy. Approximately 3,000 individuals, who had seen a mental health

professional, completed questions about their treatment. Global questions asked about how much they were helped, whether the problem for which they sought treatment improved, and the degree to which they were satisfied with their treatment. Thus, this was not a one-item global scale but multiple global items designed to characterize the experience of receiving psychotherapy. The results showed that people were generally very satisfied with their treatment and that they were helped. The different treatments that participants received did not make a difference in the results of the survey. Overall, the results could be interpreted as a glowing report of psychotherapy for a host of problems that people bring to treatment, and some psychologists took this view (e.g., Seligman, 1995) but others did not (e.g., Jacobson & Christensen, 1996).

11.3.3: More Information on Issues and Considerations

Prior to looking how the results "came out" (what participants felt in their global ratings) and before any interpretation can be made, we ought to look at the measurement instrument. We want to know from the instrument whether the results will allow us to reach conclusions either way. It would be very difficult to make the case that the ratings reflect effectiveness in light of the absence of validity data. This is not mere skepticism for its own sake. We have developed the steps for validating measures precisely to protect against drawing simple conclusions without basis. For example, the global items in the *Consumer Reports* survey might, when validated, reduce to measures of a completely different construct (e.g., how much one liked one's therapist, whether symptoms improved spontaneously). Global ratings, as any other measure, are meaningful, but that meaning requires evidence.

Global ratings raise significant problems, two of which are particularly salient:

1. The generality of the items fosters conclusions that are also likely to be general. That is, there is little precision in what is being asked. The format of a global rating usually does not permit sufficient variation (i.e., a wide range of scores from multiple items known to measure the construct) to identify differences (e.g., treatments) if they exist. Consider as an alternative for a large-scale evaluation of therapy, use of a well-developed self-report scale (e.g., MMPI-2) with multiple scales and subscales that have been thoroughly validated. If all of the treatments showed no differences on such a measure that might be more interpretable than the results from global ratings. The reason is that we already know the well-developed scale can differentiate populations, clinical problems, and status of individuals who vary in their psychological conditions.
2. Global ratings, like the one in *Consumer Reports*, often are homemade. Homemade measures are commonly used, especially in various magazines that provide questionnaires and surveys to test one's sex IQ, quality of one's partner, how good are you as a friend, and so on. These are designed to be entertaining (although methodologists jump out of their basement windows when they see them). There are no data that attest to the construct validity of the individual scales. Sometimes the magazines even report the answers to the questions and have large numbers of participants. Yet, if we do not know what the scale measures, interpretation of data from 2 to 200,000 subjects is not very different. The problem with such scales, surveys, and questionnaires from magazines is that rely on face validity, i.e., they seem reasonable to persons who invent them, to those who answer them, and those who read the results about them. Yet, there are rarely data that show the measures are valid, i.e., actually reflect the constructs of interest or indeed reliable (e.g., would show high test–retest reliability).

11.4: Projective Measures

11.4 Review the properties of projective measures in clinical research

Projective measures or techniques, as they are sometimes called, refer to a specific class of assessments that attempt to reveal underlying motives, processes, styles, themes, personality, and other psychological process.

11.4.1: Characteristics

These characteristics of projective measures are assessed indirectly. Clients are provided with an ambiguous task where they are free to respond with minimal situational cues or constraints. The ambiguity of the cues and minimization of stimulus material allow the client to freely "project" onto the situation important processes within his or her own personality.

"Projective" can be contrasted with "objective" in terms of measurement (McGrath & Carroll, 2012). As already mentioned, objective measures present various items (e.g., self-report inventories, diagnostic interviews) to which client responds. The questions or items and the answers are objective in the sense of clear, replicable, explicit, fixed, and so on.

> "Projective" presents ambiguous information and allows the clients to respond with their own answers without necessarily fixing or restricting the response alternatives (e.g., multiple choice).

Among the advantages of this approach is to move away from the ways in which self-report might be distorted

(e.g., by selecting "appropriate" answers among the answer alternative of objective measures). Yes, responses to projective measures can easily be distorted if the client in fact takes the view that something tricky is in the measure and they ought to be guarded and careful in responding.

There are many projective measures that differ according to the responses required of the subject, the type of stimuli presented, the manner in which content or style of responding is interpreted, the purposes of the test, and other factors. Among the most commonly used are the Rorschach and Thematic Apperception Test, which serve as a useful frame of reference (Groth-Marnat, 2009). These tests present stimuli to the participant and consist of inkblot designs or ambiguous drawings, respectively. The participant is required to interpret what he or she sees.

> The stimuli are ambiguous so that they can be interpreted by the client (but also by the clinician) in an indefinite number of ways.

The purpose of making the stimuli ambiguous is to examine the material or content the subject produces. Given the ambiguous stimuli, this material is considered to be a product of the individual's personality and reflect unconscious processes, underlying themes and motives, and conflicts.

Understanding projective techniques is facilitated by the old joke about the psychologist who administered several ink blot cards (the Rorschach test) after another to a client. The client was asked to respond to each card as the psychologist asked, "What do you see here?" When the client finished, the next card was presented, and so on for several cards. To each card, the client reported seeing some lurid, all-too-vividly described sexual scene that was somewhere between bad taste and oddly deviant. Of course, all of the ink blots were undecipherable (ambiguous) shapes. At the end of the testing, the psychologist finally offers her opinion and says to the client, "It seems as if you have a problem or issues with sex." The client responds, "This is not MY problem; you're the one showing all the dirty pictures." As this little story reveals, the goal of projective measures is to allow clients to reveal (project) their own thoughts, experiences, and motivations. Presumably, the less strong, clear, and "objective" the stimulus material, the more likely that the responses will be generated by processes that reflect how the client perceives, experiences, characterizes, and captures the world. Think of the client as a painter staring at a white canvas—how ambiguous and what to paint? Actually, one card from the Thematic Apperception Test *is* a blank card to which the client is asked to respond.

Responses to projective techniques are considered to be traceable to content themes and perceptual processes that unify and organize personality. Content domains, such as how the individual handles sexual or aggressive impulses, relates to authority, or expresses need for achievement as well as stylistic or coping methods such as expressing affect and managing needs, are inferred.

> Interpretations provided by the subject usually are condensed to reflect a small number of themes or processes.

Performance on projective tests has been viewed as a way to provide insights on the inner workings and organization of personality. Indeed, in conveying this point, some projective techniques (e.g., Rorschach) are considered to reflect a method to evaluate perceptual and cognitive process rather than a test per se. The measures provide broad themes, styles of coping, attitudes, and other general facets of personality.

Projective tests have been studied extensively (e.g., reliability and validity) along with various techniques and codes to score the narrative responses clients provide (e.g., Exner & Erdberg, 2005). There are debates about what various measures assess. Yet, there are validity data showing, for example, that Rorschach performance relates to distorted thinking, intelligence, effort or engagement in a task, prognosis in therapy and dependence as a personality characteristic (McGrath & Carroll, 2012). The tests are used in personality research and occasionally in clinical work in psychology. Projective measures have also been used in sports, business, and marketing where there is interest in identifying motivation, impediments to success (e.g., athletes, employees), or preferences (e.g., of consumers and in reactions to new products).

11.4.2: Issues and Considerations

Projective measures have received considerable attention in personality assessment. Their use and popularity have waxed and waned over the last 60 or so years due in part to their association with a particular theoretical approach toward the nature of personality. Developments and current topics in central interest in psychoanalytic theory (e.g., object relations) and methods of scoring diverse scales have in accelerated research on projective techniques (Tuber, 2012). Nevertheless, use of the measures is generally restricted within clinical psychology, as I mentioned, and also not taught very often in undergraduate or graduate training.

The diminished interest and use of the measures can be traced to the following factors:

- The theories connected originally with projective techniques no longer dominate clinical psychology and psychiatry. Projective measures were originally associated with and some currently adhere to psychoanalytic and psychodynamic models that explain human functioning in terms of underlying unconscious processes.

Unconscious processes are still studied extensively, but these are quite different from those posed by psychoanalysis. The "new unconscious," as it is sometimes called, refers to factors of which individuals are largely unaware in the environment (e.g., smells, sights) but that still have impact on their behavior and brain processes (Hassin, Uleman, & Bargh, 2005).

- Many projective measures traditionally have relied heavily upon interpretations and inferences of the examining psychologist. These interpretations often have been shown to be inconsistent across examiners, which has led researchers to question the basis for making judgments about personality. Scoring methods of many projective methods are somewhat cumbersome and complex, and major scoring methods have been subject to criticism (see McGrath & Carroll, 2012). The scientific evidence in support of many of the scoring methods has been challenged and rebutted (see Hibbard, 2003; Lilienfeld, Wood, & Garb, 2000).
- Other modalities of assessment have proliferated, and the methods and their foci capture current attention. As one example, neuroimaging as discussed later includes many techniques for elaborating brain activation and structural and functional properties associated with key clinical topics (e.g., psychiatric disorders, emotional states, risk for some mental or physical health outcome, and changes in therapy). Also, many clinical psychology studies focus on core psychological processes (e.g., emotion, cognition, perception) associated with clinical dysfunction (e.g., bipolar disorder, autism spectrum disorder) and lab-based measures (e.g., eye-tracking) and computerized assessment (respond to stimuli on a touch screen).

11.4.3: More Information on Issues and Considerations

Projective measures often are cumbersome to administer and score and do not assess many of the critical foci of contemporary research directly.

Thus when compared to other types of measures, such as self-report inventories, direct observations, methods of neuroimaging, or computerized assessment of cognitive functioning, projective techniques are markedly less frequently employed.

Below are listed a few additional factors that may account for diminished interest and use of projective measures:

- Even on its home turf (e.g., understanding personality) advances from objective measures have shown the benefits and progress for objective rather than projective measures. Personality includes enduring characteristics within the individual and how these influence and are influenced by interactions with the environment. Research has greatly elaborated dominant personality characteristics empirically. For one example, the most well-studied measure of personality is the Neuroticism-Extraversion-Openness Inventory (NEO-I), which assesses five personality characteristics (called the Big 5), including three characteristics named in the title of the measure plus two others (agreeableness and conscientiousness). This is an objective measure (self-report or other report) that has been revised and updated periodically (McCrae & Costa, 2010). The measure has been extensively used and validated in many different contexts in cross-sectional and longitudinal studies across children, adolescents, and adults and across many ethnic groups and countries. One recent study, for example, included well more than 1,200,000 subjects, across a wide age range (10–65 years), and 5 English-speaking countries (e.g., Soto, John, Gosling, & Potter, 2011). It is not clear whether more difficult to administer and interpret projective tests can add appreciably to other measures and provide an increment in theoretically or empirically important or clinically useful information.
- Long ago psychiatric diagnosis drew on intrapsychic explanations of various disorders. That has changed over 30 years ago. Currently psychiatric diagnostic categories focus on mental disorders from a descriptive standpoint and emphasize them as brain diseases (American Psychiatric Association, 2013; Cuthbert & Insel, 2013; World Health Organization, 2010). Decades ago, these same or very similar disorders were considered from the standpoint of putative intrapsychic and psychodynamic processes, and extracting these with projective techniques seemed more reasonable. Absence of data and the difficulty in supporting such an approach led to large revisions in describing disorders rather than imbuing them with unsupported causal and motivational mechanisms. Also, current diagnostic research has shifted the attention to neural and genetic markers of disorders and as well as core psychological processes (e.g., cognition, memory, perception).

Projective measures can raise their own obstacles that preclude them from being adopted casually into an assessment battery in a study. Thus, if the investigator would like to assess aggression, symptoms, or stress and wishes to choose multiple methods to operationalize the construct, projective tests are not the usual choice. Investigators are more likely to select measures that are more convenient to administer and score. Notwithstanding these considerations, projective techniques have occupied a very special place in clinical assessment. The full range of clinical topics including "normal" functioning of personality, characteristics

of different diagnostic groups, personality and human performance, and other areas can be evaluated from the standpoint of intrapsychic processes. Elaboration of the content areas of the field as well as development of new tests and scoring methods has made projective assessment an area of work in its own right. The future is uncertain perhaps because the techniques are unlikely to be taught in most graduate training programs in clinical psychology.

11.5: Direct Observations of Behavior

11.5 Examine the utility of direct observations of behavior in clinical research

Direct observation, as the name suggests, consists of measures that assess behavior of interest by looking at what the client actually does.

11.5.1: Characteristics

The overt behaviors noted by direct observation may be sampled from how the client performs in everyday situations or in situations that are designed explicitly to reveal specific responses. Thus, the resulting responses provide direct samples of the relevant behaviors rather than more indirect indices such as:

- Self-report measures
- Global ratings
- Projective tests

For example, we would like to know when or how often one engages in social interaction with others and texts other people. These can be assessed directly in everyday life (e.g., smartphones and apps). Even without automated measures in everyday settings, direct measures of overt behavior are of interest in evaluating interpersonal (e.g., marital) communication, sexual dysfunction, social or dating skills, enuresis, tics, stuttering, insomnia, and verbalizations of hallucinations and delusions. The fact that these problems include behavioral components does not in any way deny that other modalities of assessment are important or relevant. A key tenet of assessment is that rarely is one measure or assessment modality sufficient to evaluate a construct.

Yet, as a modality of assessment, direct observations operationalize problems in terms of non-questionnaire performance and often on samples of behavior from everyday life. Consider the contrast with self-report measures. Self-report measures have individuals say what they do (verbal behavior) by checking off responses. We already know from psychology that saying what one does is only correlated with and not the same as what one actually does and often that correlation is not very high. Direct assessment is designed to get at or more closely at what one actually does.

> Direct observation of behavior often is not quite as simple as it sounds ("Let's just observe what people do"). Behavior is a stream of actions and rarely provides the clear data we would need. Usually codes need to be developed that define what will be counted and precisely how the behavior of interest will be defined.

For example, if we want to evaluate disruptive behavior of children in a classroom or social interaction of elderly in an assisted-living setting, we will need definitions of what is and is not a disruptive behavior and what is and is not social interaction.

The codes define the units and how they will be observed. For example, for direct observation of "studying of a college student," we might count the frequency or numerical occurrences of well-defined actions (e.g., how many pages were actually read while studying), the duration (e.g., how long someone studied), latency (how long from when a student returned to his room before he began to study), or whether behavior (sitting and staring at an open book) occurred or did not occur in a given interval (e.g., 15-minute intervals from 10:00 p.m. to 12:00 a.m.). There are many options for coding when direct observations are made (Bakeman & Quera, 2012; Kazdin, 2011).

> No matter how the behavior is observed, it is essential to be sure the observations are obtained reliably. That is, there must be consistency (and little or no error) in actually coding the behavior.

Loose definitions and poor training of observers lead to data with error, and one threat to data-evaluation validity is unreliability of the measure. Stated in more disaster terms, there may have been a "real" effect and our prediction was perfectly on target. Yet, we found no differences between groups or over time for a given group because the error in the measure was too great.

Direct observations can be conducted under a variety of circumstances and in different ways. Perhaps the most obvious is conducting direct observation in the clients' natural environment (e.g., at home, at school, in the community). Sampling behavior under conditions of the natural environment or conditions resembling these is designed to assess the behavior of interest directly to diminish concerns about external validity of the findings, i.e., whether the results generalize to everyday life. Also, novel relations can be observed that are not otherwise evident.

For example, a study evaluated the relation of anger and interpersonal violence in relationships (Elkins, Moore, McNulty, Kivisto, & Handsel, 2013). Participants (college students involved in a relationship) completed several

paper-and-pencil measures, but the electronic survey is pertinent to this discussion. Participants completed a survey provided by computer (called electronic or e-diary) and answered a variety of questions daily about concrete negative relationship experiences including physical assault (e.g., grabbing, hitting, throwing something), psychological aggression (e.g., insulting, yelling, threatening), and sexual coercion (e.g., used threats or physical force to have sex, insisted on sex when the partner did not want to) over a 2-month period. Participants merely reported whether or not the specific actions had occurred each day. Similarly in another project, a mobile phone was used to assess depression of individuals as they functioned in everyday life (Morris et al., 2010). Throughout the day, adults were prompted on their phone and completed measures related to mood. They could also call up cognitively based interventions from their mobile phone as desired.

> A survey or response to mobile phone prompts completed everyday merely looks like another form of self-report rather than direct observation. It is self-report, but the distinction is made because participants report on specific behaviors or states and often in real time. They are, as it were, observers of their own behaviors. Sometimes these observations and reporting of one's own behavior in the natural environment are referred to as *experience sampling* to emphasize its focus on direct assessment in natural settings and in real time (see Santangelo, Ebner-Priemer, & Trull, 2013).

The examples I have provided are, or at least will soon be, very modest and early-stage applications of technology. We already know about smartphones, tablets, fitness monitors worn like wrist watches, and many "apps" that now permit assessment of biological processes, experience, socialization, and more. For example, one can program a smartphone to prompt individuals during the day to report on affective states and stress (Solzbacher, Böttger, Memmesheimer, Mussgay, & Rüddel, 2007). Feedback was automatically provided to help cope with these states. More generally, self-report has merged with direct observation by having individuals keep daily records about specific experiences or actions that occur throughout the day over a several day period.

11.5.2: More Information on Characteristics

Direct observations often are conducted in laboratory settings under more convenient and standard conditions than provided by natural settings. Also, laboratory conditions often permit more detailed and in-depth evaluation because the assessments can be readily recorded (e.g., videotaped for later scoring), evaluated by multiple observers (e.g., behind a one-way mirror), and use special equipment either for the subjects or for the observers (e.g., laptops, presentation of stimulus material that prompts behavior).

> Even of greater significance, laboratory arrangements can hold constant factors that could vary enormously in the natural environment.

For example, an interaction pattern, referred to as expressed emotion (EE), consists of how family members feel about and interact with each other. EE encompasses a pattern in which members tend to be critical of and hostile toward each other. EE is rather important in relation to treatment because it predicts the likelihood of relapse among patients treated for affective disorder and schizophrenia (e.g., Hooley, 2007). The extent to which family members are critical of and hostile toward each other could be studied in the home. The home conditions might be standardized in some way (e.g., no young children in the room, no incoming or outgoing phone calls, no use of weapons while observers are conducting their observations). Yet, EE is more readily assessed under standardized conditions in the laboratory where an interview of family members is provided and taped and later evaluated for comments that define EE. Also, in the laboratory, one can more readily evaluate interrater agreement to ensure that the responses were reliably assessed.

Another well-known instance of direct behavioral assessment under laboratory conditions has focused on self-control of pre-school children and specifically their ability to delay gratification (see Mischel & Ayduk, 2002; Mischel et al., 2011). In the now-famous marshmallow test, children are brought into a room one at a time and instructed by a research assistant that when the assistant returns the child can have two marshmallows. There is one marshmallow on a plate in front of the child, and the child is told she can eat that, but will get two marshmallows if she can wait until the assistant returns. The assistant leaves, and the child is sitting in a chair at a table with a plate in front with one marshmallow and can be left up to 20 minutes before the assistant returns. The entire sequence is videotaped to observe how children respond (for one of many videos, see https://www.youtube.com/watch?v=QX_oy9614HQ). Some children eat the marshmallow; some nibble at it; some wait, and there are other variations. What makes this interesting is to see individuals struggle, yield, and use various self-control strategies to resist the temptation to eat the one marshmallow. Those who eat the one marshmallow are considered to have less self-control than those who wait to obtain the two marshmallows.

This is a good example of a behavioral measure under laboratory conditions for a few reasons:

1. This is a direct measure of behavior.
2. The key concept "self-control" and observational coding system provide well-defined and replicable measures (e.g., eating the marshmallow and latency to eating).

3. The situation is standardized (held constant) in a laboratory setting, so one can see how different children respond under identical circumstances.
4. The measure conveys the challenge of any home-made measure that has all the wonderful features I just noted.

Does the measure get at anything important beyond eating marshmallows in a laboratory setting? That is, does the measure have any type of validity (e.g., concurrent or predictive)? Perhaps the measure is just cute and does not relate to anything very important.

This behavioral test, developed in the late 1960s and early 1970s, was shown in follow-up studies (40 years later) to predict social, cognitive, and mental health outcomes over the life course in adulthood (Mischel et al., 2011). For example, those with greater self-control early in life on the marshmallow test, when reassessed decades later, showed higher educational achievement, higher sense of self-worth, better ability to cope with stress, lower rates of illicit drug (cocaine/crack) use, and fewer symptoms of clinical dysfunction. Self-control and delay of gratification now are well-researched areas with different measures in use and they support and expand findings that early self-control predicts functioning in multiple domains (e.g., educational, psychiatric disorder, financial problems in adulthood, socialization) (e.g., Moffitt et al., 2011). Research also has moved to understand the potential cognitive and brain mechanisms involved in self-control (e.g., Tabibnia et al., 2011).

> Role-play tasks that simulate situations are often used to provide data for observations in the laboratory and can be especially useful if the responses of interest are low rate and unlikely to be easily observed in everyday life.

For example, role-play was used to measure how women (undergraduate students) responded to sexual coercion. Women were presented with various descriptions of sexual threats and coercion, and their reactions (e.g., negative affect, refusal) were measured directly in response to the present situations (Jouriles, Simpson Rowe, McDonald, Platt, & Gomez, 2011). Role-play is a direct sample of behavior and clearly provides measures different from self-report about how one might or would respond to situations.

Simulated situations are occasionally introduced into the natural environment to assess behaviors that otherwise would be difficult to observe. For example, an intervention program designed to train young children what to do when they encountered a handgun (Miltenberger et al., 2004). Assessment consisted of leaving the children alone in a situation with a disabled gun, videotaping what the children did, and then scoring whether various appropriate behaviors occurred (e.g., leaving the gun alone, seeking an adult). Training focused on developing the appropriate behavior, and the effects were assessed on the simulated measure after the intervention and 5 months later.

I have provided a sample of conditions in which behaviors are directly assessed but cannot begin to detail the range of options and opportunities. Direct assessment of response domains (e.g., emotions, cognitions) and clinical functioning (e.g., symptoms, stress) is likely to accelerate in light of technological advances. From an assessment standpoint, we have strong interest in sampling how an individual functions in everyday experience and technological developments afford more opportunities to do that. Advances in assessment in real time are changing delivery of psychological treatments as well because the data can be used to generate self-help coping strategies, cognitive behavior therapy interventions, and text messages for support.

11.5.3: Issues and Considerations

Because the behaviors of interest are observed directly, the measures seemingly are straightforward indexes of the problems or response patterns. There are multiple qualifiers to note. To begin, direct samples of behavior are not necessarily representative samples of what behaviors are "really" like. Yes overt behavior seems to get at "real" behavior because it skips self- and other-report filtering (via perception, memory) of behavior. Also, such reports are a bit removed from what the individual actually does. (In some situations, verbal behavior [what people say whether or not it reflects other behavior] may be the "behavior" of interest.) Yet, decisions regarding what to observe could restrict interpretation and generality of the measure. The codes that were made to observe behavior have some arbitrary definitional features to permit observation. What is and is not "emotional abuse" in a relationship may not capture all of the behaviors that could be involved or give sufficient weight to those behaviors that many would find especially abusive.

In addition, the observations may not represent the behavior based on when the observations are obtained. It is possible that the sampled behaviors or periods of time when assessment is conducted do not accurately portray the client's performance at other times. If the periods of observation samples (e.g., 1 hour of observation per day) are to represent all of the potentially available observation periods (e.g., all waking hours), assessment methods need to ensure that there are no differences that occur across the available periods of assessment. This can be accomplished by randomly selecting periods throughout the day for observation. Although this is not feasible for most behaviors from practical considerations, it would seem to resolve the problem of obtaining a direct and representative sample of behavior. More important perhaps than randomness of the period in which behavior is assessed are the

conditions in which assessments are conducted. Individuals are aware of assessment and indeed in the usual situation must be as a matter of ethical obligations of the investigator. Performance may change when individuals are aware they are being studied.

> As more and more assessments can be completed from mobile devices in use (e.g., smartphones, wrist watch, and bracelet monitoring devices) or super high-tech work in progress (e.g., in one's clothing, car, home), many of the concerns about unrepresentativeness of samples of behavior are reduced.

One can assess behavior all day (e.g., waking hours), or large segments of the day, random samples throughout the day, or "oversample" problem periods (e.g., eating when one returns to one's dorm, apartment, or home at the end of the day). Also, many assessments can be recorded automatically out of the awareness of individuals. This is not difficult to program with an application and smartphone. Thus, more studies can use assessment in real time and capture information as the person functions in everyday life. Awareness of assessment tends to attenuate with time and when assessment is automated as part of one's smartphone, it will not stand out as special or outside of routine (e.g., carrying a smartphone around).

As noted earlier, many direct observations are made in contrived situations in the laboratory. Yet, performance in contrived situations may differ considerably from what would be reflected in everyday life. Marital interaction and communication in a laboratory may reflect dysfunction but still not resemble very closely the nature of the interactions in everyday life in the privacy of one's own home. Participants may be aware of the special assessment arrangement and respond differently as a result (e.g., show less intense conflict and no physical abuse). Simulated situations are not inherently limited. However, direct observations cannot be assumed to be valid, i.e., relate to performance in other settings any more than other types of measures (e.g., self-report inventories). Validity evidence is needed to draw conclusions about the generality of the measures to the extent that the conditions of measurement differ from those of behaviors in everyday life, as illustrated by the example of the marshmallow test.

On balance, direct observations provide a unique focus that extends the method of evaluation beyond the more familiar and commonly used self-report scales and inventories. The special feature that has accelerated the assessment of overt behavior is the use of technology. Increased sophistication of hardware, software, sensing devices, automated storage and transmission of data in real time, and feedback based on real-time data alter assessment as well as the ability to intervene (e.g., with coping strategies, useful feedback). These assessments are sometimes direct sampling of behavior, i.e., what a person does. In many instances, affect, cognition, and biological processes will be detected directly, and the assessment is more direct samples of something but not necessarily "behavior" or actions. Also, as we have seen, some of the assessments are self-report in real time in which individuals report on what they are doing or what they are feeling or thinking. This self-report in real time is not necessarily behavior either but is different from self-report on a questionnaire that reflects a one-shot assessment outside of the context of everyday moment-to-moment experience.

11.6: Psychobiological Measures

11.6 Express the properties, pros, and cons of psychological measures in clinical research

Psychobiological measures refer to assessment techniques designed to examine biological substrates and correlates of affect, cognition, and behavior and the links between biological processes and psychological constructs.

11.6.1: Characteristics

There have been enormous advances in the available psychobiological measures and the scope of domains within psychology to which they are applied. The measures encompass many different types of functions (e.g., arousal of the autonomic system), systems (e.g., cardiovascular, gastrointestinal, neurological), and levels of analysis (e.g., microelectrode physiology that permits analysis of the response of individual neurons in the brain and brain imaging in response to tasks and activities in human and nonhuman animal research).

Measures are obtained in many different ways:

- Connecting subjects to noninvasive apparatus (e.g., to assess respiration, heart rate, blood pressure, electrical activity of the brain)
- Connecting subjects to apparatus that are a little more invasive (e.g., to assess sexual arousal, mentioned later in the chapter)
- Sampling saliva or drawing blood to assay a range of biological metabolites

Measures within this domain are quite different from types of measures we have discussed and involve many different methods only a few of which can be sampled here.

Within psychological research, many interventions target areas that are related to psychological states (e.g., mood, anxiety, attentiveness, vigilance, stress, arousal, and others).

Some of the more familiar biological measures include heart or pulse rate, blood pressure, skin temperature, blood volume, muscle tension, and electrical activity of the brain.

Biological measures directly reflect many domains of interest related to physical and mental health in clinical psychological research.

Psychobiological measures are the primary measures in many areas of clinical research. For example, a great deal of clinical research focuses on the onset, course, treatment, and prevention of the use of drugs (e.g., marijuana), alcohol, or tobacco.

In one program to treat cigarette smoking, the primary outcome measure was the level of carbon monoxide (CO) in the blood (Glenn & Dallery, 2007). Individuals breathed into a CO monitor over the course of the study. This measure has been well studied, so one can evaluate levels (parts of CO per million) that are known to reflect abstinence from cigarette smoking. Similarly, in a study designed to decrease marijuana dependence, an oral swab test was used to evaluate drug use (Twohig, Shoenberger, & Hayes, 2007). The test requires placing a special pad between the lower cheek and gum for 2–5 minutes. The results indicate whether marijuana was used within the past 3 days. These are just samples of biological measures substance use, which is an area central to clinical psychology.

The use of biological measures of substance use is more readily familiar in the context of amateur and professional sports and conveys the obvious benefits of this method of assessment. The World Anti-Doping Agency (www.wada-ama.org/en/) is the leading international agency that campaigns for doping-free sports. Testing is one part of the work behind the overall effort to permit better detection of substance use. One could readily use a self-report measure to assess substance use by asking world-class athletes right before their event in the Olympic events whether they have used any illicit substance. There might be a group self-report measure used where a psychologist or one of the judges shouts over a megaphone to all the Olympic weight lifters, "Before we start, would all of those who use steroids step forward." One might expect "socially desirable" response set where participants tried to place themselves in a good light by staying put. Psychobiological measures are obviously essential. There are assay techniques (e.g., from blood serum and urine) that provide validated even if not perfect or foolproof measures of substance use and give finer-grained information (e.g., how much or how recent the use) (e.g., Lund et al., 2013). For example, one commonly used technique is mass spectrometry which examines physical, chemical, or biological properties of compounds at the molecular level. In general, the testing methods become increasingly sensitive and improved overtime and can better detect use of illicit substances.

Indeed, urine and blood samples can be held for several years (e.g., up to 8) after the sporting event to take advantages of emerging and yet to be developed methods to test for illicit substances.

Substances that cannot be detected now might be readily detectable soon.

Psychophysiological measures have been used extensively in evaluation of sexual arousal and sexual dysfunction.

Some of the research has focused on evaluating or altering sexual arousal in persons who experience arousal in the presence of socially inappropriate and censured stimuli (e.g., exhibitionistic, sadistic, masochistic stimuli, or stimuli involving children, animals, or inanimate objects). Sexual stimuli can be presented in the actual situation or by computer or video to determine whether they arouse the clients. Arousal to the stimuli can be assessed directly by looking at blood volume changes in the penis (penile plethysmography) or lining of the vagina (vaginal photoplethysmography). For example, a penile plethysmograph measures blood volume based on a band around the penis that registers increases in the diameter (e.g., Reyes et al., 2006). This is a well-studied and validated measure of sexual arousal and sexual preference. Such assessment does not replace or obviate the need for a self-report assessment of arousal, but rather illustrates direct assessment of the physiological aspects of arousal.

Measures of physiological arousal and reactivity are used extensively in clinical psychological research in part because they relate to core topics (e.g., anxiety, stress, pain) and because technological advances have facilitated assessment (e.g., portable and noninvasive measures that do not require high-level technical maintenance) and data collection (e.g., automated scoring and conversion to a database). For example, studies of response to stress may use self-report measures, but also are likely to use such measures as heart rate to convey through more direct measures the extent to which stress has been induced.

Measures such as heart rate, skin conductance, respiration, blood pressure, and many other such measures can be obtained by connecting subjects to apparatus while they are engaging in experimental tasks.

Outside of the context of anxiety, other commonly used measures focus on muscle tension (e.g., electromyographic [EMG] responses) and electrical activity of the brain (e.g., electroencephalographic [EEG] responses). Technological advances have made such assessments easier to complete and more user-friendly for the subjects (e.g., portable, small equipment, as opposed to ominous looking wires connected to several places in one's body).

As I have noted, smartphone and related devices now can assess an increasing array of psychologically and biologically relevant constructs (e.g., socialization, stress).

One biochemical measure used frequently is cortisol level. This is a good example because it is a noninvasive measure assessed directly from samples of saliva, although occasionally blood samples are used. Cortisol levels and changes reflect critical neuroendocrine functioning (the limbic-hypothalamic-pituitary-adrenal axis).

> Cortisol often is used to assess degree of stress and stress reduction in response to intervention.

Also, cortisol has been assessed extensively to assess the extent to which individuals are stressed in an experimental arrangement, or to delineate subtypes of individuals (high vs. low reactivity), and to evaluate possible mechanisms in response to various activities (e.g., Walker et al., 2013; Yehuda & Seckl, 2011). Neuroendocrine functioning and changes in functioning, as assessed by cortisol, have been implicated in a wide variety of conditions (e.g., child abuse and maltreatment, gambling, and psychiatric disorders) that are topics of research in clinical psychology. Arguably the most prominent and increasingly used measures that qualify as psychobiological are those based on neuroimaging and other measures of neural processes.

11.6.2: More Information on Characteristics

In psychology, behavioral, social, and cognitive neuroscience is a huge area of work. The pervasive use of the methods is evident in less familiar increasingly used terms as clinical neuroscience and cultural neuroscience (e.g., Chambers, Garavan, & Bellgrove, 2009; Kitayama & Park, 2010).

Diverse measures also are used heavily in clinical psychology and psychiatry in an effort to evaluate neural processes associated with various psychiatric disorders or states and also to characterize changes that result from interventions (e.g., medication, psychosocial treatment). Depending on the specific measures, structure, function, and processing can be examined to in observational (e.g., case control comparing diagnostic groups or "healthy" controls) and experimental studies (e.g., intervening to alter mood states) and with human and nonhuman animals. Table 11.4 provides a sample of some of the more commonly used measures at this time and what they assess. Yet, assessing neural processes is an area of rapid development, and the sample can only present some of the highlights. Obviously, imaging techniques require quite special equipment, facilities, training, and collaborations (with nonpsychologists) and hence have not been standard fare in assessment batteries in most programs of clinical research. Yet an increasing number of clinical psychology faculty have programs of research that are based on neuroimaging, and training programs hiring new faculty often have special interest in hiring someone whose work draws on a range of brain-related assessment techniques.

Table 11.4: Selected Neuroimaging and Assessment Techniques

Technique	What Is Measured
Functional Magnetic Resonance Imaging (fMRI)	Measures the changes in blood oxygenation and flow that occur in response to neural activity. Greater activity in a region of the brain utilizes more oxygen and the increased blood flow to achieve that is measured.
Computed Tomography (CT)	Devises a picture of the brain based on absorption of tissues of X-rays. The individual lies inside a scanner while X-ray beams passing through the head help reveal features of the brain based on differential absorption of the rays.
Positron Emission Tomography (PET)	Radioactive material in small (trace) amounts is used to map processes of the brain. As the radioactive material decays, a subatomic particle (positron) is emitted and this is detected to measure high areas of activity of the brain.
Single-Photon Emission Computed Tomography (SPECT)	As in PET, it uses radioactive tracer that is taken up by the brain quickly and reflects cerebral blood flow and provides images of active regions of the brain.
Electroencephalography (EEG)	Measures electrical activity of the brain by recording from electrodes placed on the scalp. The resulting traces reflect electrical signals from many neurons near the site of the electrodes.
Magnetoencephalography (MEG)	Maps the brain activity through recordings of magnetic fields produced by electrical currents that naturally occur in the brain.

Among the neuroimaging measures, the most familiar is fMRI (functional magnetic resonance imaging), which permits the investigator to identify areas of the brain that are activated when individuals are given a task to perform. The nature of the task can call on different psychological abilities (e.g., memory, problem solving, efforts at emotional regulation) or emotions (e.g., love, disgust) and much more. From activity that is evident, one can hypothesize neurological processes that might be involved. Activation of brain centers is a just a beginning but helps to integrate other findings about critical areas and how they operate.

Imaging techniques provide opportunities to identify and distinguish different psychiatric disorders (e.g., depression, schizophrenia), subtypes of disorders, individuals who have and have not recovered from disorders, and changes over time and how these relate to symptom change (e.g., Hamilton et al., 2012; Masdeu, 2011). This is not merely measuring the brain activation of different clinical populations. Rather one can begin to explore a variety

of psychological processes (e.g., memory, processing information) from the multiple brain sites that are activated, and how these vary and are similar across diverse disorders. From that one can develop a model of neurological underpinnings and correlates of clinical dysfunction.

Apart from evaluating diagnoses and their associated features, neuroimaging has been used in several interesting ways related to the topics of clinical psychology (e.g., Frewen, Dozois, & Lanius, 2008; Linden, 2006; Porto et al., 2009; Quidé, Witteveen, El-Hage, Veltman, & Olff, 2012; Roffman, Marci, Glick, Dougherty, & Rauch, 2005). Studies using various imaging techniques have been used to:

- Measure experimentally induced or provoked symptoms (e.g., sadness manipulations in healthy samples; trauma stimuli in PTSD patients) to demonstrate experimentally brain areas implicated in dysfunction.
- Show "normalization" of neurological structures, function, and activity after psychotherapy therapy is completed and symptoms of a disorder have remitted.
- Show similarities and differences in specific brain processes altered by different interventions (e.g., medication, psychotherapy) for a given disorder (e.g., major depression).
- Show some similarities in what brain processes are altered by the same intervention (e.g., cognitive behavior therapy) as applied to different disorders (e.g., obsessive compulsive disorders, depression).

In short, neuroimaging plays an increasing and increasingly diverse role in clinical psychological studies and the psychological processes that may characterize, contribute to, predict, or result from psychiatric disorders.

Some of this work involves imaging of nonhuman (as well as human) animals because states can be induced in animals to mimic psychological states of interest in human functioning (e.g., stress, anxiety, and depression). Also, one can apply interventions (e.g., medication and exercise—both evidence-based treatments for depression) to see how they operate in nonhuman animals. The scope of research and the extent to which imaging techniques are used are remarkable. For example, a recent review identified 224 studies of neuroimaging in human and nonhuman animals that focused on the neurological effects of cannabis use on mental health (Batalla et al., 2013). (You may have not been aware that nonhuman animals who use marijuana have mental health issues—not sure that they do but experimental studies can induce both marijuana use and psychological states that mimic psychological symptoms.)

More generally, there are extraordinary advances being made in imaging that can be used with either or both human and nonhuman animals. Less frequently used and newer techniques will augment our evaluation at multiple levels (e.g., electroencephalographic and magnetoencephalographic methods, event-related potential, near-infrared spectroscopy, transcranial magnetic stimulation, diffusion tensor imaging, two-photon microscopy, spectral imaging, fluorescence lifetime microscopy, and fluorescence anisotropy analysis; e.g., Perkel, 2013; Sporns, 2010). Whole brain and individual neuron analyses will permit watching how processes (e.g., learning) unfold also in real time (e.g., Underwood, 2013). The varied measures give different and complementary pictures and elaborate the structural and functional connectivity of the brain. As an assessment method, the vast array of techniques can add greatly to other modalities of assessment as relevant to a given hypothesis.

Psychobiological measures have obvious benefits in assessment and have figured prominently in many areas of clinical research, only a few of which I could mention. Psychobiological measures often are the central to the primary goal of the study, but even when they are not they can be very useful to incorporate. The measures, as all measures, have their own sources of artifact and error, but they are less subject to some of the common artifacts that seem to plague many other measures.

For example, response patterns such as socially desirable responding and acquiescence do not seem relevant or as relevant when monitoring such measures as heart rate, blood pressure, and respiration. Also, voluntary alteration of responses to psychobiological measures in light of demands of the experiment situation is likely to be less than the alteration on self-report or behavioral test measures. For these reasons, psychobiological measures often have been regarded as direct measures to circumvent many sources of artifact and bias present in other modalities of assessment. Of course, psychobiological measures have their own sources of problems, artifact, and bias. And debates exist about units of analyses, variation of software used to code neuroimages, and how brain activity relates to psychological processes to mention a few (e.g., Brown, 2012; Harley, 2004). Yet, the multiple ways now available to map neural processes and the increasingly fine-grained analyses will make connections of all sorts that will eventually elaborate mechanisms and processes of how experience translates to biological functioning and vice versa (Nelson & Sheridan, 2011).

In many ways, neuroimaging and other techniques are not merely assessment methods to get at constructs assessed in other ways. They have opened up new content areas and more fine-grained analyses. This is happening in many areas and sciences where technology (e.g., assessment of genes and the brain) and methods of analyses (e.g., social network analyses, "big data" sets) are revealing new phenomena to which we previously did not have access and startling relations that open new fields. For example, the microbes in our body (e.g., friendly bacteria—or so they seem) look like they play a critical role in learning,

hormone regulation, immune system, disease, and cancer treatment and no doubt yet to be revealed other areas (see Human Microbiome Project at http://commonfund.nih.gov/hmp/). Advances in assessment are not just advances in assessment. They can give us access to realities that are present but to which we have not had access.

11.6.3: Issues and Considerations

Psychobiological measures is a term I have used for presentation purposes. Yet the measures include remarkably different foci (e.g., from blood, breath, saliva, and neurotransmitters, gene processes) and arguably could be assigned to many different categories. They share in a broad abstract feature of drawing on biological processes or indices to inform topics of interest within clinical psychology. We have known for some time that measures of physiological states (e.g., stress) do not always go together, and so there is no one way to characterize a state or experience (e.g., stress, pain). Some common indices may emerge, but not all measures of a given state or experience correlate well or show the same pattern among different individuals. For example, a set of measures within a given system (e.g., heart rate, blood pressure, and blood volume as measures of the cardiovascular system) and across systems (e.g., measures of cardiovascular functioning, respiration, skin resistance) in response to specific events (e.g., in the laboratory) may not be related in a consistent fashion for different subjects. This is important to note to convey that the specific measure used can very much influence the conclusions that are reached as one measure could show a given pattern less clear from another measure seemingly getting at the same or closely related construct (e.g., arousal). Similarly, we also know that similar outcomes or presentations of how things look might well have quite different underpinnings. For example, symptoms of depression can derive from a psychiatric disorder but also from neurogenerative (e.g., Alzheimer's) disease. Neuroimaging can help make the diagnoses because studies have shown different processes are involved to produce similar symptoms (e.g., Masdeu, 2011). These points underscore thoughtful consideration of the measures that are selected to evaluate the system of interest.

There are practical and mundane considerations that can influence the use of psychobiological measures:

- Psychobiological recording often requires rather expensive equipment, particularly if multiple response systems are monitored simultaneously.
- Someone in the lab usually is needed to maintain, repair, and calibrate the equipment and ensure that interpretable data are obtained.
- In many cases (e.g., neuroimaging), this goes well beyond the availability of a laboratory technician and requires strong collaborative arrangements with individuals in other fields (e.g., diagnostic imaging, physics).
- Use of the equipment (e.g., time on the magnet for neuroimaging) too can be quite costly from a few hundred to several thousand dollars per hour depending on the type of scan and the analyses that will be drawn from that. Clearly, the expense is prohibitive for some measures. Yet, for others measures that are not so difficult to obtain (e.g., blood samples, saliva to measure cortisol or to obtain DNA for genetic analyses), procedures to maintain the samples and to ensure their proper analyses are obviously critical.

Artifacts unique to particular assessment methods can influence responsiveness on measures. Movements of the subject, changes in respiration, electrical interference from adjacent equipment, and demands of the situation may enter into the responses of subjects who are connected to various devices. Whether the potential sources of artifact occur are in part a function of the particular measures and the nature of the recording system. For example, inadvertent or intentional changes in respiration on the part of the participant can affect heart rate data and can introduce artifacts. Such influences can be readily controlled or addressed by monitoring systems that might mediate changes in the response of interest or by ruling out the possibility of involvement in a specific system by removing its influence (as in the case of animals given curare so that skeletal responses cannot alter heart rate). Psychobiological measures provide unique information and levels of analysis in relation to the available assessment modalities. The measures continue to develop in two directions:

1. Higher resolution and finer-grained methods of assessing brain processes and functions no doubt will continue to emerge, and these require continued advances in hardware and software. The advances have permitted and will continue to elaborate mechanisms, processes, and substrates of more complex and dynamic biological functions.
2. More physiological measures are likely to be available that permit wider use beyond well-equipped laboratories.

   More portable, less expensive, and user-friendly measures also have increased (e.g., caps that can be worn to assess EEG activity or sleep patterns, automated blood pressure cuffs).

Thus, many measures have become more practical and less expensive and can be more easily integrated into assessment batteries. No doubt more sophisticated physiological measures will find their way into apps and Web-based assessments, a topic to which we now turn.

11.7: Computerized, Technology-Based, and Web-Based Assessment

11.7 Scrutinize how computerized, technology-based, and web-based assessment has helped in clinical research

This category of assessment modality is unique and might be divided in different ways and even absorbed into other categories. The reason is that the use of computers and technology represents novel ways to administer measures of other modalities (e.g., self-report) but also has helped to generate novel ways of assessing functioning (e.g., implicit assessment methods). The categorization is not as important as conveying the options for assessment.

11.7.1: Characteristics

A few decades ago, computers came to be used as part of psychological assessment. Material could be presented in an automated way, on screen, or even by voice and elicit responses via self-report (by endorsing response alternatives) and by task performance (e.g., responding to cues or decision making and touch screen).

> *Computerized assessment came to refer to the use of computers and automated collection of information as well as scoring and evaluating that information.*

The topic goes well beyond the present discussion and encompasses many alternative test formats, the use of computers for different facets of testing (e.g., administration, interpretation), client reactions to computers, ethical issues (e.g., privacy of Web-based assessment), and others. Also, the range of electronic devices that can be used for assessment well beyond a "computer" is enormous. The focus here is on administration of measures via computer and encompasses the Internet and Web-based assessments as well.

As hard to imagine as it is, computers were the sole advanced technology just a few decades ago, and it made sense to refer to computerized assessment.

> Technology and advances in all sorts of devices, including smartphones, tablets, and other monitoring devices with specific purposes, can assess all sorts of domains and using modalities already discussed (e.g., self-report, psychobiological processes). Computerized assessment still is a heavily used category, plays a role in clinical research, and warrants highlighting.

With computerized assessment, an individual is presented with a task on a computer or merely as part of a usual experimental arrangement, and responds to the computer (touch screen monitor, key board) to convey responses. As an example, computerized psychiatric diagnosis instruments where the individual answers questions presented directly on screen (e.g., Computerized Diagnostic Interview Schedule, Quick Diagnostic Interview Schedule [Quick DIS]) have been around for over 25 years and continue in active use in research (e.g., Brown et al., 2010; Westermeyer & Canive, 2012). Among the advantages can be brevity of administration. As soon as criteria are met for a diagnosis, the remaining pertinent questions are asked or the measure moves to the next set of symptoms. Such a measure is particularly useful when a large number of interviews need to be administered and personnel costs would be high. For example, in one study of the Quick DIS, more than 1,000 medical and surgical hospitalized patients (males in various veterans medical centers) were studied to examine the extent to which medical disorders are associated with psychiatric disorders (Booth, Blow, & Cook, 1998). Use of computerized diagnostic assessment made this feasible. Incidentally, the results showed that almost half (47%) of patients met criteria for at least one psychiatric disorder over the course of their lives. This is in keeping with other research that has shown that individuals with chronic medical conditions have high rates of psychiatric disorders.

Web-based assessment is an important extension of technologically based assessment. Many studies include assessment on the Web in someway (e.g., via MTurk and Qualtrics).[2] Yet, there are broader ways in which this can be done as well. As an illustration, programs to reduce and prevent bullying in the schools often use multiple interventions involving administrators, teachers and staff, parents, and students. Among the goals is to alter the climate of the school in addition to other strategies beyond the scope of the present discussion. A key to the program is assessment of bullying and its many forms to obtain data that clarify the specific types of bullying as well as the prevalence of the problem. Assessment is done in an ongoing way and serves as feedback for the effectiveness of the program. For example, in the United States, one state (Maryland) conducted a district-wide school anti-bullying program. Assessments were Web-based to collect ongoing information on bullying and included more than 25,000 students, 2,000 staff, and 800 parents and involved 116 public schools (Bradshaw, Debnam, Martin, & Gill, 2006). Individuals could log in and provide data on diverse facets of bullying. The information could evaluate school-wide prevention programs, and look at bullying at individual schools, types of bullying, and other details.

> Web-based assessment provides opportunities for administration on a large scale and allows completion of the measure under circumstances that are convenient (e.g., from one's home or work).

The different conditions under which individuals complete the measure can introduce variability into the assessment process, a definite trade-off. The example with bullying is use of the Web format for presentation of self-report and hence might be classified as an efficient self-report measure.

11.7.2: More Information on Characteristics

Computer-based assessment goes beyond reformatting other modalities of assessment and can be used to provide novel ways of measuring characteristics. A familiar example that has been applied widely in psychological research is the Implicit Attitude Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Greenwald, Poehlman, Uhlmann, & Banaji, 2009). The goal is to measure attitudes that are not immediately accessible in a person's awareness but reflect views about how a person might feel or his or her attitude toward some concept (e.g., views toward a given ethnic group or oneself). Participants are presented with concepts on a computer screen (e.g., black, white). One concept appears at the left of the screen; the other at the right. Then the attribute or adjective appears (e.g., smart, unpleasant), and the participant sorts the word into one of the categories by pressing a key for the left or right category. There are variants of the task within the IAT, which requires the participant to respond to categories and attributes to which they are connected in the presentation on the screen. From the tasks, one can identify the stronger associations between categories and attributes. A category (e.g., black) and attribute that go together yield faster responses from the participant than those that do not.

> The underlying view is that the measure bypasses conscious thinking and gets at implicit, unconscious, or automatic reactions (associations).

The measure has been used heavily to evaluate stereotypes, racial and gender bias, and many more characteristics. In some instances, the measure has been used to "show" that most of us are prejudice in light of associations we make once consciousness is bypassed. This has been quite controversial because of many other interpretations that might explain response latencies. The assessment issue is whether the measure is valid and relates to other criteria and there is evidence of concurrent, predictive, and incremental validity (Greenwald et al., 2009) but controversy remains about the strength of the relation in tests of validity (i.e., low correlations) and whether the IAT format improves among more explicit and direct measures (Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Previously I mentioned an example where direct (self-report) assessment of marital satisfaction among newlyweds showed a zero correlation with an implicit attitude measures of one's partner (McNulty, Olson, Meltzer, & Shaffer, 2013). What made the findings interesting is that over the next 4 years, the self-report ("conscious") measure did not predict marital satisfaction in later assessments but the implicit measure did.

There are demonstrations that attest to the utility of the IAT in the context of clinical psychological research. For example, suicidal thoughts are usually evaluated by self-report, but the self-report format allows one to distort and withhold information in reporting such thoughts. Perhaps the format of the IAT could contribute. In a study with three groups of adolescents (attempters, ideators, and nonsuicidal controls), the IAT was used to measure implicit associations of self-injury and oneself (Nock & Banaji, 2007). Performance in the IAT was able to discriminate (classify) among the groups and also predicted future (6 months later) suicidalideation. The measure added an increment in predicting suicidality (incremental validity) over and above other known predictors (e.g., mood and substance use disorder, total number of psychiatric disorders, past and present psychiatric dysfunction, prior suicide ideation and attempt). Thus, the value of the measure was quite clear. The IAT format has lent itself to assessing many different constructs in quite diverse contexts (Wittenbrink & Schwarz, 2007). This is one illustration of a much larger assessment format in which computerized assessment presents material and participants respond by pressing the screen or keyboard to measure a wide range of domains (e.g., cognitive functions, neuropsychological deficits).

Computer-based assessment is one narrow option that may soon look primitive as a way us using technology. On the near horizon are a variety of sensing and monitoring devices that can be worn to assess diverse facets of physical health (e.g., blood pressure, vision, alcohol levels).

> Physical and mental health-related sensors of special interest for psychology would be sensors to evaluate cigarette smoking (number of puffs while smoking), drinking alcohol, consuming drugs, experiencing stress, and engaging in daily conversation episodes.

Additional measures are being developed to help with smoking relapse by sensing vulnerable moments that might lead to smoking and triggering an immediate intervention to help (e.g., Ertin et al., 2011; Kumar, Nilsen, Pavel, & Srivastava, 2013). In addition, personal physical and mental health care is likely to be improved because of such assessments and the integration of assessments not only in cell phones, but also in one's car and home. Measures are increasingly feasible for large-scale use because one of the goals of this research has been to improve health care with better and more pervasive monitoring of health within and cross culturally. (For a preview of some coming assessment attractions in health care, see www.youtube.com/watch?v=DP7jbExNbJw#at=1126.)

11.7.3: Issues and Considerations

When computerized assessment first emerged, a key question was whether the assessment was valid and yielded results that were comparable to those obtained with the usual in person clinical assessment. No statement can be made that applies to all measures, constructs, and samples. However, many studies have been completed in the context of social, emotional, and behavioral problems, academic functioning, screening and diagnosis, and across multiple disciplines and applications and have shown that results are comparable to noncomputerized assessment. That is, correlations are high between the standard way of administered the measure (e.g., live interviewer) and computerized administration (e.g., Gottschalk et al., 2000; Hallfors, Khatapoush, Kadushin, Watson, & Saxe, 2000; Randall, Sireci, Li, & Kaira, 2012). Evaluation and use of some measures are particularly well developed (e.g., MMPI-2) with many studies on administration, scale analyses, and interpretation (see Nichols, 2011). Also, computerized and automated or Web-based assessment now has been used in several studies to evaluate psychosocial intervention programs, so it is clear that measures can reflect therapeutic change in many contexts. Automated assessment by computer or other device no longer need to be proven as a medium of delivery, but of course any novel assessment and construct combination will invariably need to traverse the steps to establish reliability and validity.

There are a number of advantages of computerized and other technology-based assessment:

1. Presentation of items has various advantages associated spontaneity and assessment in real time. Assessment may be self-report on emotional, cognitive, or behavioral experiences in everyday life on some mobile device. Even if the questions were identical to those on a paper-and-pencil questionnaire administered in a lab, presentation in everyday life on multiple occasions during the day with the task to respond to these questions based on what is happening or how one feels right now goes well beyond a mere change in presentation of the items. Arguably, the task and measurement are qualitatively different by computerized administration in this way.
2. Computerized assessment offers efficiency and individualization of assessment. A computerized test need not merely be a fixed set of items the way a questionnaire might be. Based on client responses, other items can be called up that elaborate a particular symptom domain or diagnosis. Also, questions may be purposely skipped, i.e., not presented to the subject, because the subject's responses to those questions can be shown to be unnecessary based on prior statistical evaluation. The measure is objective and standardized in one way (e.g., presentation format and "core" items) but is also individualized in the sense of permitting the presentation of items to elaborate characteristics the client shows and to present only those items relevant to a given client in light of his or her pattern of responding to early items (e.g., Gibbons et al., 2008, 2012).
3. Other advantages to computerized assessment include:
   - More reliable administration of the measure (e.g., not skipping questions)
   - Elicitation of more information (e.g., in a given area as further details are asked or as eliciting more revealing information than a human clinical assessor would)
   - Incurring lower costs (e.g., personnel costs of test administrators or interviewers)
   - Allowing large-scale application
   - Increasing the reliability of clinical decision making (see Garg et al., 2005; Parshall, Harmes, Davey, & Pashley, 2010)

   Also, computerized measures can go well beyond merely presenting items. Video, animations, sounds, and graphics can be presented, and participants can be more actively involved in how they respond (selecting, touching the screen, respond to moving items).

   All of these latter features provide a new type of testing altogether. Often computerized assessment is more preferable to the subjects than are clinician-administered interviews. This benefit occasionally leads participants to reveal information that they would not otherwise reveal in the presence of a live examiner.

Technology-based assessments as I mentioned previously are advancing with entirely new devices to measure indices of health, including both physical and mental health (e.g., mood, stress, blood pressure) (e.g., Ertin et al., 2011; Kumar et al., 2013). Among the goals is to develop means of assessment for large-scale application so that preventive and treatment interventions can be implemented in everyday context to monitor health routinely. These would be devices one can carry like exercise monitoring devices or cell phones or are embedded in things we wear.

From the standpoint of the present chapter, computerized and technology-based assessment provides another modality of assessment. Computerization does not reduce or alter method factors (variance attributed to any particular assessment modality). The strength of assessment in a given study comes from using multiple measures of a construct and varying the methods of assessment. Computerized assessment provides an alternative with several practical and cost advantages.

Computerized assessment has been well studied and applied to many areas within psychology (e.g., neuropsychology, clinical psychology, cognitive psychology) and

many other areas of work (e.g., education, sports) (e.g., Gottschalk & Bechtel, 2009; Maarse, Mulder, Brand, & Akkerman, 2003). The benefits of computerized assessment have yet to be fully developed. One benefit not yet realized is in the context of clinical work. Routine, computerized assessment would provide low-cost information that could easily improve the data (opinions, judgment) on which clinicians rely for decision making about patient care, progress, and when to alter or end therapy. The goal is not necessarily to replace judgment but rather to provide better data (e.g., in everyday life, in real time) than what is currently available.

11.8: Unobtrusiveness Measures

11.8 Describe unobtrusiveness measures as used in clinical research

Most measures used in psychology are administered under conditions in which subjects are aware that assessment is being conducted and aware (even if only approximately) of the purpose of the measure. Awareness of measurement (obtrusiveness) can lead to changes (reactivity) in how individuals respond to the measure. Projective measures have as a goal overcoming the level of awareness of what is being measured and thereby reducing an individual's ability to distort responses. Computerized measures such as the IAT also seek to circumvent the level of awareness that measures (e.g., self-report scales) provide. In both projective techniques and IAT, as examples, individuals can indeed distort their performance and provide impressions (e.g., socially desirable responding) (e.g., Crowne & Marlowe, 1964; Cvencek, Greenwald, Brown, Gray, & Snowden, 2010; Röhner, Schröder-Abé, & Schütz, 2013). Indeed, in almost all types of measures mentioned previously, participants are aware that their performance is being assessed, whether or not they know the specific purposes or foci of the measures. Also, ethical issues and subject protections ordinarily require that individuals are aware of being assessed. There are exceptions (e.g., if the identity of the participant is not known or used, if records are assessed in which participants remain anonymous).

> Unobtrusive measures are a type of assessment that are out of awareness of the person whose behavior or other characteristics are being assessed (Kellehear, 1993; Schmidt, 2012; Webb, Campbell, Schwartz, & Sechrest, 2000).

11.8.1: Characteristics

The major techniques of unobtrusive measurement are listed in Table 11.5. The techniques include simple observation, observation in contrived situations, archival records, and physical traces and are listed in Table 11.5 for a convenient summary. In some instances, the measures are not "new" (e.g., direct observation) and could be placed in a prior type of measure I discussed. Yet, the category of unobtrusive measures is unique and helps us ponder novel ways of evaluating key constructs of interest that depart from the much more commonly used obtrusive measures (e.g., self-report questionnaires).

Table 11.5: Major Methods of Unobtrusive Measurement

Type of Measure	Definition	Examples
Simple Observation	Observing behavior in naturalistic situations in which the assessor does not intervene or intrude. The assessor is passive and does nothing to alter the normal behavior or to convey that behavior is being observed.	Observing nonverbal gestures or body distance as a study of social behavior; recording clothing individuals wear to reflect mood states.
Observations in Contrived Situations	Simple observation of behavior in naturalistic situations where the experimenter or assessor intervenes or does something to prompt certain kinds of performance. The assessor plays an active role without violating the reactivity of the situation.	Using confederates who seem to be in need of assistance to test for altruism; testing for honesty in a situation that allows for cheating.
Archival Records	Records kept for reasons other than psychological research such as institutional, demographic information, photographs, social media, or personal records.	Records of birth, marriage; institutional data, such as discharge records or patient history; documents; Facebook or texting records.
Physical Traces	Physical evidence, changes, or remnants in the environment that may stem from accumulation or wear resulting from performance.	Wear on pages to discover magazine or book passages read, marginal notes and comments on book pages; deposit of trash to study littering, graffiti to study sexual themes.

Simple Observation: Directly observing behavior as it occurs in naturalistic settings, unbeknownst to the subject, is the most obvious unobtrusive measure and could have been included in the previous discussion of direct observation. Simple observations in naturalistic settings sample behavior unaffected or less affected by the situational constraints of the laboratory and methodological characteristics of the more commonly used assessment procedures.

The fact that the observations are outside of awareness of the subjects eliminates or at least minimizes reactivity. The actual behaviors that are observed too may be

relatively subtle and ones that subjects would not suspect reflect the construct of interest.

An example that has been pursued in many psychological studies is touch, i.e., people touching each other (Hertenstein & Weiss, 2011). There are fascinating patterns to consider in relation to who touches whom, in what contexts, and how that influences subsequent interactions. Early intriguing work that stimulated work in the area was a creative evaluation of power and gender issues. One function of touching other individuals (e.g., having a hand on another person's back, putting an arm around someone's shoulder, holding someone's arm while talking to them) is to convey status or power (Henley, 1977). Higher status or more powerful individuals may be more likely to touch others than to be touched by others. If touching is a sign of power or unwitting efforts to display power, then individuals with higher status or who wish to convey that status (e.g., persons who have higher socioeconomic status, are older, male) all would be expected to touch others (i.e., their respective counterparts) more than be touched. In fact, unobtrusive observations of touching in public situations supported the prediction. Individuals who were male, older, and rated as higher in socioeconomic status more frequently touched others (females, younger individuals, persons of lower socioeconomic status, respectively) than were touched by them. The findings do not establish that touching necessarily assesses status or power. Yet the observational data supplemented questionnaire research that related touching others to dominance, status, and being placed in a position of power. Thus, direct observation adds credence to other assessment methods for evaluating social behavior. It would be useful to exhume touching as power to see if changes in the culture in relation to sex and perhaps cultural groups reflect or no longer reflect differences in unobtrusive power displays.

Simple observation is very useful because of the almost unlimited situations in everyday life that are open to scrutiny and direct tests of hypotheses. Of course, the method has potential problems, such as:

1. A problem might arise which would defeat the value of unobtrusive observation in naturalistic situations is detecting the presence of the observer qua observer. As an unobtrusive measure, the observer must not influence the situation. Usually this amounts to disguising the role of the observer, if an observer actually is required in the situation. If performance can be sampled without observers, perhaps by Web cams and hidden cameras, even less opportunity might be present to alter the nonreactivity of the situation.
2. A problem with simple observation is ensuring that the behaviors of interest occur with sufficient frequency to be useful for research purposes (e.g., differentiation of groups, data analyses). Merely watching participants in the situations of interest does not guarantee that the responses will occur. This assessment problem is very familiar for various television shows that capture the predatory behavior of tigers and lions as they hunt and devour an animal. The behavior is filmed for later editing and television viewing and is clearly much better than giving animals a self-report questionnaire. Yet, to get the footage of the behavior requires many hours because the base rate of the "desired" behaviors (hunting, killing, devouring) is low in relation to hours in the day. Also, animals are not always as successful as the television shows imply. For example, cheetahs catch their prey about 50% of the time; two lions hunting together catch their prey only about 27% of the time (http://animals.pawnation.com/cheetah-catch-prey-2381.html; www.edge.org/3rd_culture/myhrvold_lions07/myhrvold_lions07_index.html). Add to the challenge, when successful hunting does occur, it may not be easily assessed for practical reason (e.g., the photographer could not keep up with the animal, a huge rock, bush, canyon, blocked the camera). It is easy to envision that scores of camera people resting in the tall grass waiting for the right shot far outnumber the animals they are viewing. Simple observation in naturalistic situations for research has similar obstacles. The response of interest may be so infrequent as to make assessment prohibitively expensive, inefficient, or of little use. Remote web cams and automated recording no doubt can help redress practical obstacles, and this has been well exploited in monitoring communities for low base rate crimes.
3. A final problem with simple observation for research purposes pertains to the standardization of the assessment situation.

   The environmental conditions in which the response occurs may change markedly over time. Extraneous factors (e.g., presence of other individuals) may influence behavior and introduce response variability in the measure.

The net effect of this variability might be to obscure the effects of the independent variable. Simple and naturalistic observation can be influenced by uncontrolled factors that may make it difficult to assess performance in a relatively uniform fashion.

11.8.2: More Information on Characteristics

Observation in Contrived Situations: Observations in contrived situations resolve some of the problems of simple observation. Contrived situations maximize the likelihood that the response of interest will occur. Hence the problem of infrequent responses or conditions that do not precipitate the response is resolved. Also, arranging the

naturalistic situation allows for standardizing extraneous factors, and hence the data are less subject to uncontrolled influences.

The important requirement of observations in contrived situations, of course, is to control the situation while maintaining the unobtrusive conditions of assessment. This may be accomplished by utilizing an observer, experimenter, or confederate working with the observer, to help stage the conditions that are designed to evoke certain kinds of behaviors.

A prime example of contrived situations for the purposes of assessment are television programs (*Candid Camera*, *Totally Hidden Video*) that place people into situations varying in degrees of frustration.

The situations are well planned so that as each new unwitting subject enters into the situation (e.g., a cafeteria), the stimulus conditions presented to him or her are held relatively constant (e.g., someone sitting next to the subject wearing a feathered hat that keeps hitting the subject in the face while he or she is eating at the counter). The subject's behavior is recorded on film, which serves as the basis for the television program. The reactions of the participants when they are informed that they are being filmed for the show often reveal the success in hiding the contrived nature of the situation. Of course, even though the conditions are relatively natural, subjects may occasionally see through them. Even so, seeing how individuals actually do respond in such situations is likely to be quite different from what they would say on a self-report questionnaire about how they would respond.

Contrived situations represent a viable option for research. They provide the control to prompt the responses of interest and can overcome the low base rate of the behavior that might occur with simple observation. Also, it may be possible to isolate the situation (e.g., part of a shopping mall, people waiting in line) so that facets of the situation that are naturalistic do not vary too much.

Archival Records: Institutional records in schools, medical facilities, government (e.g., census data), work history, use of various social services, credit history, and the Internet (e.g., what sites we visit, for how long) provide a wealth of information about people.

The unique feature of such records is that they usually can be examined without fear that the experimenter's hypothesis or actions of the observers may influence the raw data themselves, although there are exceptions.

As an example, one study used the use of emotion words for approximately 100 million Facebook users to assess "national happiness" (Kramer, 2010). This construct was operationalized as the standardized difference between the use of positive and negative words as aggregated over time. Users were all from the United States and English speaking. The study showed that the measure correlated with self-reported satisfaction with life (convergent validity). Here is an example of using archival records to evaluate a psychological construct of interest.

Archival records have their own sources of measurement issues. One problem is the possible changes in criteria for recording certain kinds of information. For example, records of crime rate may vary over time as a function of changes in the definition of crime, sociological variables that may alter the incidence of reporting certain crimes (e.g., rape), and mundane issues such as whether there are budget cuts in an agency that affect whether or how carefully the data are gathered. The changes in the criteria for recording information (an example of instrumentation) may lead to interpretive problems regarding the "true" rates of the problem and changes over time.

A related problem is the selectivity in the information that becomes archival. For example, historical records of births are likely to omit many individuals. Before more extensive methods of recording births and population statistics came into use, many births were likely to have gone unrecorded. Those births unlikely to be recorded may have varied as a function of socioeconomic status, age, and marital status of the mother, geographical location, and race. Thus, there may be a selective deposit of the information that becomes archival for subsequent research.

Physical Traces: Physical traces consist of selective wear (erosion) or the deposit (accretion) of materials. Either the wear or deposit of materials may be considered to reflect specific forms of behavior. An excellent example of a physical trace measure was used to evaluate the long-term impact of lead exposure in school-age children.

Lead is a heavy metal to which individuals can be exposed through multiple sources, including water, air (e.g., from leaded automobile fuel exhaust), paint, and other sources. Lead leaves a physical trace by accumulating in one's bones and teeth. Ethics and Institutional Review Boards that must provide approval before research can begin tend to be a little testy when an investigator proposes removing bones from children as part of research. A creative alternative was reported decades ago in a classic study that helped to change international standards for acceptable levels of lead exposure.

In this study, collecting bones consisted of baby teeth that were normally extruded (Needleman & Bellinger, 1984). Teeth were collected from thousands of children to assess lead deposits. High and low lead exposure children groups were formed from this assessment and compared in their academic and classroom performance over a period of several years. The results indicated that relatively low doses of lead exposure are associated with hyperactivity, distractibility, lower IQ, and overall reduced school functioning in children (Needleman, Schell, Bellinger, Leviton, &

Alldred, 1990). Moreover, follow-up 11 years later showed that these impairments were maintained. This work has led to other studies (e.g., replicating the deleterious effects of lead on child behavior, nonhuman animal research locating specific sites in the brain that are deleteriously affected) and to changes in government policies regarding the control of lead levels. The use of physical traces of heavy metals in teeth is still in use along with more sophisticated ways of analyzing these traces (e.g., mass spectrometry) to evaluate whether neurotoxins are implicated in clinical dysfunction. For example, teeth, hair, urine, and blood samples have shown that heavy metals are related to autism spectrum disorder (e.g., Abdullah et al., 2012; Adams et al., 2009; Al-Farsi et al., 2013; Geier, Kern, King, Sykes, & Geier, 2012). Although this is still an area of debate (e.g., interpretation of the meaning of the differences, what toxins are involved), not all studies are finding the effects.

Potential problems with physical trace measures are:

1. Changes over time may occur as a function of the ability of certain traces to be left. For example, research on graffiti in public bathrooms has focused on differences in the frequency of inscriptions between males and females and interventions that may be effective in reducing these inscriptions (e.g., Matthews, Spears, & Ball, 2012; Mueller, Moore, Doggett, & Tingstrom, 2000). Beyond the bathroom, individuals who mark surfaces with graffiti are much more likely to engage in other criminal activity (e.g., Taylor, Marais, & Cottman, 2012). If one wished to study graffiti over time, as a physical trace, this might be difficult. Many institutions have "seen the writing on the wall" and have used surface materials that are less readily inscribed or cover marks before they accumulate. Thus, the material upon which traces are made may change over time.
2. The selective deposit of physical traces is another potential problem. Physical trace measures may be subject to some of the same limitations of archival data. It is possible for the traces to be selective and not represent the behavior of all the participants of interest. Also, physical traces may be influenced by a number of variables that determine what marks are left to evaluate and hence what data will be seen. For example, fingerprints are the example par excellence of a physical trace measure. However, they are not always available as signs of someone's presence at the scene of a crime. Individuals not interested in leaving such traces are well aware of the necessary procedures to ensure that their presence and fingerprints go unrecorded. Yet, criminals are more likely to leave samples of DNA (e.g., from hair, blood, in the case of sexual crimes, semen).
3. A final problem with physical traces is that they may become reactive. Once the trace becomes known as a measure of interest, potential subjects may become aware of this may respond accordingly.

   For example, social scientists and news reporters occasionally have a keen interest in the trash of celebrities and politicians to measure their private affairs (e.g., correspondence) and potential vices (e.g., weekly consumption of alcohol). Publicity about these practices probably has limited the types of items that are publicly discarded for trash pickup.

   Secretive, cautious, and perhaps wise celebrities alike may use other means of disposal (e.g., paper shredder, trash compactor).

11.8.3: Issues and Considerations

Unobtrusive measures have several advantages:

1. They can supplement more commonly used techniques and thereby add strength to the external validity of experimental findings. For example, unobtrusive measures of therapy outcome (e.g., hospital visits, days of work missed) would provide tremendously important information about treatment efficacy and would uniquely supplement the data obtained from the more frequently relied upon self-report questionnaires and inventories. If findings are obtained across diverse measures with different methodological features (e.g., obtrusive and unobtrusive measures), this suggests the robustness of the relation between the independent and the dependent variables.
2. Unobtrusive measures often have persuasive appeal. Such measures are often drawn from archival records in everyday life (e.g., arrest rates, doctor visits, and truancy). Research that reports such measures often is much more persuasive to consumers of research (e.g., policy makers) because the measure (rather than the construct) is of interest in its own right. For example, showing that exercise of mindfulness in well-controlled studies reduces scores on the best self-report inventories and questionnaires is not likely to be viewed by outsiders as nearly as important as showing that visits to medical doctors or lost days of work are reduced. As psychologists we are concerned primarily with validity of measures, but another entirely different dimension is credibility or perceived relevance of our measures, i.e., the likelihood that those who might profit from our work see the findings as credible in light of the assessment. Thus, adoption and dissemination of findings may be improved by supplementing more commonly used psychological measures with unobtrusive measures in which society has interest. An alternative is to

show a measure that does not look very much like everyday life (e.g., marshmallow test) in fact relates to critical outcomes (e.g., drug use, psychiatric symptoms) decades later.

3. Finally, unobtrusive measures often have practical advantages. They can be accessed without subject cooperation or consent, they can be harmless and anonymous, and often they are inexpensive to collect. These features vary with the specific measure.

Unobtrusive measures have their own problems:

1. Apart from the issues mentioned already, each of the measures must be interpreted with some caution.
2. Unlike more commonly used measures, unobtrusive measures usually undergo little validation research, so there are few assurances that they measure what the investigator wishes to measure.
3. In addition, whether the unobtrusive measure will be sufficiently sensitive to reflect the relation of interest is difficult to determine in advance.
4. Finally, there is the possibility that the measure, unless contrived, is recorded in a selective way and hence does not represent the behavior of interest.

In general, there is less collective experience with a given unobtrusive measure than with standardized measures such as questionnaires and inventories. The diverse types of reliability and validity are not readily known for most unobtrusive measures.

> Unobtrusive measures need to be corroborated with other measures in the usual way that assessment devices are validated. This can be done both by empirical research that examines the relation among different measures and by theoretical formulations that place a particular measure into a context that makes testable predictions.

Increasingly greater confidence can be placed in the measure as additional predictions are corroborated. This logic, of course, applies to any psychological measure, whether or not it is unobtrusive. Some of the unobtrusive measures can raise very special ethical issues. Research obligations to participants require that they provide informed consent regarding assessment and intervention facets of the experiment.

> Unobtrusively observing performance in everyday life and using information to which subjects have not consented violate the letter and spirit of consent.

On the other hand, archival measures and physical traces may not raise concern because they address past performance and could not threaten or jeopardize in any way the identity of the individual participants. Yet, even here one has to be cautious. It is possible that the measures will place a particular group (e.g., ethnic) or locale (e.g., geographical) that unwittingly casts a sample or population in a poor light. The identity of specific individuals is not the only concern in research, a point discussed later on the topic of ethical issues. The very nature of unobtrusive assessment (perhaps no informed consent, no immediate option to withdraw from the study) means that the investigator must be quite sensitive to possible ethical concerns when these measures are contemplated.

11.9: General Comments

11.9 Examine how a modality can be best suited for certain clinical research

This overview of major types or categories of assessment is not intended to be complete either in terms of the number of modalities available or the variations within each modality. Major options were highlighted that may be differentially relevant for an investigation depending upon the purpose and constructs of interest on the part of the investigator. Selection of a given type of assessment might be dictated by theoretical predictions, the focus of the investigation (e.g., emotions, actions, cognitions), or outcomes one is seeking in the case of preventive or treatment research.

> *The discussion has focused on the type of assessment free from the content areas of clinical, counseling, and related areas of psychology.*

Often the measures are dictated by the content area and the interests they inherently reflect. For example, within clinical psychology a great deal of research focuses on neuropsychological assessment. The area considers the diagnosis and evaluation of functioning and damage to the brain as, for example, associated with injury, psychological dysfunction, medical illness, and aging. A variety of specific measures and tasks are routinely included to assess intellectual skills, sensation, memory, speech perception, tactile discrimination, and other domains (see Lezak, Howieson, Bigler, & Tranel, 2012). Many measures are regarded as standard to address the range of questions that neuropsychological assessment requires. For other areas of research in clinical psychology as well, one might identify measures and modalities in frequent use. The issue for our discussion was options for selecting dependent variables more generally.

In some studies, it may be difficult to discern precisely why one modality of assessment was selected rather than another. Yet the description of the purpose of the research should provide a rationale as to why a particular modality has been selected. Within that modality, it is desirable to convey further the rationale for selecting a particular measure. In most cases where the rationale is not explicit, there

may be extensive evidence attesting to the utility, reliability, and validity of the assessment technique. In other cases, many options might be available and the decision appears arbitrary or perhaps looks slavishly following some other study that itself did not make a strong case for the type of measure or specific instrument.

> Specific hypotheses about the constructs that constitute the dependent measures may dictate not only the modality of assessment (e.g., psychobiological measures) but also the particular measure within the modality (e.g., heart rate rather than skin conductance).

For a given research or clinical purpose, one modality may be more well suited than another because it reflects the construct and level of analysis of interest (e.g., report of significant others rather than or in addition to self-report; behavioral measures for samples of everyday interactions). However, one type of measure is not inherently superior to another. The investigator's purpose or concern over a particular source of bias or artifact may dictate which modality of assessment and measurement devices within a given modality will be appropriate.

As I have noted, in general it is valuable to use multiple measures and different types of measures. Each type of measure includes different sources of bias and potential limitations. There is no single measure that overcomes all of the problems that arise in assessment. Indeed, the measures are complementary. Selecting several different measures, each with different sorts of problems, increases confidence that the response dimension (construct) of interest in fact is being assessed. Using separate measures can help distinguish those responses that may be due to methodological idiosyncrasies of a given assessment device from systematic changes in the construct or domain of interest.

Summary and Conclusions: Assessment: Types of Measure and Their Use

Several types of measures were covered that are used in clinical psychological research. These included objective and projective measures; direct observations; psychobiological measures; computerized, technology-based, and Web-based assessment; and unobtrusive measures. Specific measures were occasionally used as examples without an effort to enumerate the enormous range of options within a type of measure. Obtaining specific measures for a study requires integration of issues and selection criteria noted in this chapter.

In general, it is useful to rely upon multiple measures rather than a single measure because:

1. Constructs of interest (e.g., clinical problems, personality, social functioning) tend to be multifaceted and no single measure that can be expected to address all of the components.
2. Performance may vary as a function of the assessment method and devices used.
3. An individual's standing on a particular dimension or construct is partially determined by the method of assessment.

It is useful to demonstrate that changes in the construct of interest (e.g., anxiety) are not restricted to only one method of assessment. Essentially, demonstrations relying upon multiple assessment techniques strengthen the confidence that can be placed in the relationship between independent and dependent variables. As I noted, often there is one measure that is of primary if not sole interest (e.g., subjective views of happiness, survival) and so multiple measures may not be as critical.

For each type of measure, I discussed various strengths and limitations. A reason to utilize different types of measures is to ensure that not one type of limitation (e.g., socially desirable responding, transparency of the measure) characterizes the full assessment battery. Thus, self-report may be the most straightforward measure, but the demonstration might be strengthened by measuring the same construct in yet another way.

Critical Thinking Questions

1. What are objective measures? Why is the term "objective" misleading?
2. What is reactivity of measurement, and why should we care about it?
3. What might be a good "app" (make up a hypothetical one) to assess a psychological construct or behavior using a smartphone, smart watch or bracelet, or other equivalent device.

Chapter 11 Quiz: Assessment: Types of Measure and Their Use

Chapter 12
Special Topics of Assessment

Learning Objectives

12.1 Outline experimental manipulation with respect to the independent variable and subsequent effects on the subjects

12.2 Determine the different types of manipulations as carried out in experimental research

12.3 Describe two situations when the manipulation checks provide useful insights

12.4 Analyze four possible situations involving the manipulation check and the dependent measure with respect to interpretive issues

12.5 Analyze some of the main point areas that arise due to ambiguity on how and when to use manipulation checks

12.6 Examine how the interventions are evaluated with respect to treatment, prevention, education, and enrichment programs

12.7 Express how regular and periodic evaluation of client progress is carried out during the course of treatment

The main facet of assessment is selecting measures to serve as dependent variables in one's study. We have covered criteria to consider when selecting measures (e.g., various types of reliability and validity) and many of the types of measures (e.g., objective measures, Web-based). There are other facets of assessment:

1. Experiments in which some manipulation is provided to participants and consists of assessing the impact of the experimental manipulation. That is, we manipulated some experience (e.g., what the participant was exposed to, told, saw, or was led to believe), and now we want to check whether our manipulation "took," which means was grasped, perceived, or experienced by the participant. Assessing the impact of the experimental manipulation is a check on what we have done to see if our manipulation worked.
2. Evaluation of interventions and is appropriate for measures in clinical psychology, psychiatry, education, counseling, and areas where we are evaluating client change and want the client to benefit directly in some way. Changes on our dependent measures may not provide information that shows the changes made in the clients are important or make a difference in some palpable way. Clinical significance is the term used to reflect such changes in the context of treatments for psychological dysfunction. "Applied significance" and "practical significance" are other terms and can be applied more broadly (e.g., education, counseling) beyond psychotherapeutic interventions.
3. Ongoing assessment in a project, and here too the focus is on intervention work. In research on treatment, usually measures are given before and after the intervention. There are special benefits in adding interim measures that are evaluated over the course of treatment. Two situations in which the ongoing assessment is especially useful are in clinical work with patients seen in clinical practice and in research designed to evaluate mediators of change.

All three topics comprise the focus of this chapter.

12.1: Assessing the Impact of the Experimental Manipulation

12.1 Outline experimental manipulation with respect to the independent variable and subsequent effects on the subjects

In an experiment, the independent variable is manipulated by:

- Providing a particular condition to one group and omitting it from another group
- Providing varying degrees of a given condition to different groups
- Presenting entirely distinct conditions to groups

Providing an experience to induce mood in the experiment (e.g., sadness vs. happiness) is one example. Great care is required to ensure that the variable or condition is manipulated as intended and that the manipulation delivery is consistent across participants within a group (e.g., all subjects intended to receive the manipulation in fact do get it). Careful control and administration of the manipulation are required for interpretation of the findings (construct validity) and for a sensitive evaluation of the manipulation (data evaluation validity).

12.1.1: Checking on the Experimental Manipulation

The hypothesis of interest in the investigation is based upon the assumption that the independent variable was effectively implemented. It is extremely useful to check whether the independent variable, experimental manipulation, or intervention was implemented as intended.

> *Providing a check on the manipulation refers to assessing the independent variable and its effects on the subjects in ways that are separate from evaluating whether the manipulation has impact on the dependent variables of interest.*

Assessment of the independent variable or a manipulation check is distinguished both procedurally and conceptually from dependent variables included in the study. The check on the independent variable assesses whether the conditions of interest to the investigator were altered or provided to the subjects. This may mean merely that the stimulus was presented as intended or that the intervention was received or perceived by the subject. For example, in a mood induction study, a check on the independent variable would be achieved by actually asking subjects to rate their mood (e.g., self-report, selecting a face from a screen that best captures how they feel) or using some other measure (e.g., facial expressions, casual comments to a confederate). These checks on the independent variable are quite separate from the dependent measures (e.g., performance on some task that was intended to be altered by mood state).

> In one sense, the best check on the effects of an independent variable is the dependent measure because the change in the independent variable is intended to alter the dependent measure.

If the predicted results of an experiment are obtained, assessment of the independent variable to ensure that it has had the intended effect on the subject may not seem to be essential. Presumably, the independent variable accounted for the results, barring obvious threats to internal and construct validity. Even so, it is possible that the change on the dependent variables occurred for reasons other than the manipulation check (e.g., novelty, expectations, and demand characteristics) and maybe the manipulation was not even picked up very well by most of the participants. Checking the extent to which the independent variable is effectively manipulated provides information that can greatly illuminate the findings. There are hidden caveats too we shall uncover.

12.2: Types of Manipulations

12.2 Determine the different types of manipulations as carried out in experimental research

The way in which the success of the experimental manipulation can be assessed varies as a function of the type of manipulation or independent variable.

12.2.1: Variations of Information

In many experiments, the manipulation refers to different information given to subjects across experimental conditions. The initial question to be answered for the check on the success of the manipulation is whether the information was in fact delivered by a research assistant or experimenter. Assume that the information was provided to the participants as intended. The check on the manipulation is whether participants received, attended to, and believed the information.

Typically, a manipulation check consists of providing subjects with a questionnaire immediately after hearing the rationale or viewing some brief vignette to convey the information. For example, the independent variable might consist of telling subjects about some aspect in the news, some personality characteristics of a hypothetical individual who is later evaluated as part of the study, or arguments designed to change beliefs about some healthful or

unhealthful practice. Presumably different groups in the study will vary in the information that is presented. A questionnaire might be administered after the instructions were administered or later in the experiment to assess whether the subject heard or grasped the information. If subjects respond to alternatives that reflect what they were told in their respective experimental conditions, the investigator could be more confident that the independent variable was manipulated as intended.

When the experimental manipulation relies upon information, self-report questionnaires are frequently used to assess the success of the manipulation. A few questions might be all that are needed.

> These questions might be in a true-false, multiple-choice, or open-ended (essay question) format. It is useful to include a true-false or multiple-choice format in most cases so that each question can be easily scored and answered.

Open-ended questions might be used, such as "What information did you learn from the experimenter?" or "What did the experimenter say when you began the experiment?" These questions seem useful and maybe even clever because they do not reveal the purpose of the experiment or give away the correct answers as readily as true-false or multiple-choice questions. Indeed, the multiple-choice questions may give away that there was a manipulation and what it might have been (sensitization effect) and lead individuals to realize there even was a manipulation. Open-ended questions are better in one sense, but they can be very difficult to score. Special codes are required, and it is more likely that subjects did not get the point of the questions. Many subjects do not answer them, reply with only one or two words, or elaborate extended discussions that miss the point of interest to the investigator.

In general, when the independent variable involves variation of information to the subject, the manipulation check is relatively straightforward. There usually is a check to ensure that the experimenter delivered the information, that it was received or perceived by the subjects, and that experimental groups are distinguishable on a measure that assesses that information. Self-report measures are commonly used because they are readily adaptable to the experiment by merely constructing examination-type questions.

12.2.2: Variations in Subject Tasks and Experience

Many manipulations consist of having subjects do something, engage in a particular task, actually carry out the instructions, or experience a particular state. The research question of interest is likely to be whether a certain task facilitates or hinders some outcome. For example, subjects may play a game in which the stakes or the level of competition is varied or engage in some cognitive tasks. Checking on the manipulation would consist of evaluating whether the subjects reported or responded in such a way on the task as to reflect their understanding of the manipulation. Here the check on the manipulation may be performance on the task to convey the participants understood what to do and actually did that. The manipulation check for task variables assesses what subjects do rather than what they know or even say they did (self-report). Even so, answering questions (self-report) about specific tasks and activities that were performed may be used and is better than not assessing the manipulation at all.

> The manipulation may consist of more than a task. For example, instructions, activities, or tasks may be designed to induce a particular mood or emotional state in the subject.

Exposure to the experimental manipulation or task alone is insufficient as a manipulation check. For example, the purpose of an experiment may be to induce high levels of euphoria or similar states in some of the subjects and moderate or low levels in other subjects. The investigator wishes the subjects to experience something in a particular way and the experimental test depends on achieving this state. The manipulation check refers to the participant's state or performance on the task or activity and goes beyond merely being exposed to the manipulation. Maybe I blinked, tuned out, did a quick tweet, or checked my e-mail and missed the crucial phases of the manipulation—as an experimenter we might want to identify those individuals who somehow did not experience the manipulation as intended.

If a self-report measure were used to the experimental manipulation, items would be included to allow participants to report the extent to which they experience euphoria. Items might have participants rate on a 5-point scale how euphoric they feel (1 = not at all euphoric, 3 = moderately euphoric, 5 = very euphoric). The investigator could infer with some degree of confidence that the independent variable was successfully manipulated if groups differed in the extent of euphoria on their ratings according to the respective conditions to which they were assigned. Whether the independent variable has impact on the dependent measures is another matter, but at least the investigator would know that the manipulation was implemented as intended and registered with the subjects.

12.2.3: Variation of Intervention Conditions

Many interventions in clinical research consist of varying the conditions to which subjects are exposed. Primary examples of this type of manipulation would be exposing subjects to:

- Different therapy
- Prevention

- Counseling
- Educational interventions
- Remedial interventions

In these cases, the manipulation is implemented or carried out by the therapists or trainers.

In the simplest case, one group receives the intervention and the other group does not (treatment vs. no treatment or a waiting-list control). Specification of the experimental manipulation consists of how well treatment was delivered. Obviously, the implied hypothesis in that treatment when conducted appropriately and as intended is likely to produce greater change than no treatment. Of primary interest is an evaluation of the extent to which treatment was conducted as intended, a concept referred to as *treatment integrity* or *treatment fidelity*.

Treatment integrity is important to assess and relevant whether or not treatment outcome differences are evident and whether one treatment was more effective than another. A study comparing two or more treatments, for example, may show that both treatments "worked" but were no different in their outcomes. A pattern of no difference might result from a failure to implement one or both of the treatments faithfully or diffusion of treatments. For example, one review noted that short-term psychoanalytic treatment was as effective as other therapies (Leichsenring, Rabung, & Leibing, 2004). Yet, the conclusion was rejected by others because of the inattention to treatment integrity. The original study included no assurances treatments were carried out as intended (Bhar & Beck, 2009). A "no difference" finding might be readily be explained by lapses in integrity, but we need the data on integrity to make that more or less plausible.

Large variation in how individual treatments are carried out across patients within a given condition (within-group variability or error in the statistical analysis) and blending or mixing of treatment conditions that ought to be distinct (diffusion of treatment and reduction of between-group differences needed for statistical significance) could readily lead to no differences. Ensuring treatment integrity can help avoid these pitfalls. Even when two treatments differ, it is important to rule out the possibility that the differences are due to variations of integrity with which each was conducted. One treatment, perhaps because of its complexity or novelty, may be more subject to procedural degradation and appear less effective because it was less faithfully rendered. Thus, integrity of treatment is relevant in any outcome study, independently of the specific pattern of results.

> The breakdown of treatment integrity is one of the greatest dangers in intervention research. The danger stems from the many opportunities for interventions to breakdown.

12.2.4: Additional Information on Variation of Intervention Conditions

Unlike laboratory experiments, interventions usually consist of multiple sessions, so integrity means correct implementation on many occasions and sometimes with many different procedures that vary session by session. The danger also stems from the fact that most well-controlled outcome studies do not measure treatment integrity. In fact one evaluation examined treatment research from several premier journals (e.g., *Journal of Consulting and Clinical Psychology, Archives of General Psychiatry*) spanning over 200 different treatments in over 140 studies; only 3.5% intervention studies assessed treatment integrity (Perepletchikova, Treat, Kazdin, 2007). This means that we cannot be assured that interventions were implemented as intended or that we know that all of the individual components of treatment were actually delivered. Most investigators recognize the importance of assessing treatment integrity, but report many barriers (e.g., cost, time, and labor demands of doing this extra assessment, lack of guidelines on who to assess integrity) (Perepletchikova, Hilt, Chereji, & Kazdin, 2009).

There are several steps that can be performed to address treatment integrity (Leichsenring et al., 2011; Perepletchikova & Kazdin, 2005):

1. The criteria, procedures, tasks, and therapist and patient characteristics that define the treatment ought to be specified as well as possible. Many investigators have described their treatments in manual form, which includes written materials to guide the therapist in the procedures, techniques, topics, themes, therapeutic maneuvers, and activities, as readily seen from searching "psychotherapy manuals" on an Internet search engine (e.g., Google Scholar) or book order sites (e.g., Amazon). When treatment is explicitly described, it is easier to develop guidelines to decide when a session or treatment was or was not delivered as intended and what level or type of departures is considered tolerable within the study.
2. Therapists can be trained carefully to carry out the techniques. It is useful to specify the requisite skills for delivering treatment and to provide training experiences to develop these skills (e.g., role-play, practice cases with sessions that are videotaped for feedback and further training).

   > Videotapes of "good" or prototypical sessions can be used to convey the style and to provide guidelines regarding how that style is likely to be achieved.

   Years of experience in providing a treatment, often a criterion espoused in clinical work, is not an adequate criterion for stating or assuming that therapy was

administered well or with integrity. Experience alone does not ensure proficiency in adhering to a specific technique or set of techniques. Indeed years of experience without feedback or supervision can lead to systematic lapses or looser execution of procedures over time. Providing special and uniform training experiences for the therapists (experimenters, trainers) is useful and can have important implications for how faithfully treatment is likely to be rendered.

3. When treatment has begun, it is valuable to provide continued case supervision. Listening to or viewing tapes of selected sessions, meeting regularly with therapists to provide feedback, and similar monitoring procedures may reduce therapist drift (departure) from the desired practices.

   If there are multiple therapists, group feedback and supervision sessions are especially valuable to help retain homogeneity in how treatment is implemented across a heterogeneous group of people.

Whether treatment has been carried out as intended can only be evaluated definitively after the treatment has been completed. This evaluation requires measuring the implementation of treatment. Audio or videotapes of selected treatment sessions from each condition can be examined. Codes for therapist and/or patient behaviors or other specific facets of the sessions can operationalize important features of treatment and help decide whether treatment was conducted as intended. Checklists can be used to assess whether specific discrete tasks were performed; ratings can be used to assess more qualitative and stylistic features, if these too are relevant to integrity. Research assistants or staff not involved in treatment can observe recordings of the session and complete the measures designed to evaluate integrity.

Treatment integrity is not an all-or-none matter. Hence, it is useful to identify what a faithful rendition of each treatment is and what departures fall within an acceptable range. For some variables, decision rules may be arbitrary, but making them explicit facilitates interpretation of the results. For example, to consider a relatively simple characteristic, treatment may consist of 12 sessions of individual psychotherapy. The investigator may specify that "receiving treatment" or an "adequate test of treatment" consists of any instance in which the client received 75% or more or so many weeks of the sessions. For other variables, particularly those within-session procedures that distinguish different treatments, specification of criteria that define an acceptable range may be more difficult. In some cases, the presence of select processes (e.g., discarding irrational beliefs, improving one's self-concept) might be sufficient; in other cases, a particular level of various processes (e.g., anxiety or arousal) might be required to denote that treatment has been adequately provided.

12.3: Utility of Checking the Manipulation

12.3 Describe two situations when the manipulation checks provide useful insights

Data showing that the independent variable was manipulated as intended increase the confidence that can be placed on the basis for the results. Two situations are worth highlighting because manipulation checks provide particularly useful information, namely:

- When the experiment produced no significant differences between groups
- When experimental conditions are especially important to keep distinct

12.3.1: No Differences between Groups

If the predicted results of an experiment are not obtained on the dependent measures and, in fact, no significant group differences are evident, assessment of the independent variable may prove to be remarkably helpful in interpreting the results. As an example, consider an experiment that provides a manipulation to two groups that differ only in what the subjects are told. The goal is to vary expectancies that they will probably be stressed by what they watch (e.g., brief videos of surgery where a significant amount of blood is evident). The goal is to see if expectations for high or low stress will influence the next step of having individuals distract themselves (think of something else) as a technique to regulate their emotional reaction (i.e., reduce stress). (Both groups receive the same videos—these are two different stressful videos.) Two are used to avoid the stimulus sampling construct validity threat:

- One half of subjects in each group receive one video
- Other half receive the other video

The high expectancy group is told that the video is really horribly stressful and they might not be able to watch it all. The low expectancy group is told the video is not very stressful and pretty routine surgery. The hypothesis is that the high expectancy group will be able to cope better when asked to distract themselves right after the video and are then asked to complete a measure of stress and happiness.

> Suppose the results show no differences—expectancies did not make a difference at all. What can be said about the impact of the instruction/expectancy manipulation on stress, happiness, and love of methodology measures?

It is very important to ask whether the independent variable was manipulated adequately so that the different

instructional sets were salient to the subjects. Certainly, we would want to know whether subjects heard, knew, or believed the expectancies. If the subjects did not hear or attend to the crucial instructions, then the results of the study would be viewed differently from the situation in which the subjects fully heard and believed the instructions. If the subjects had not perceived the instructions, then the hypothesis under study was not really tested. That is, "objectively" we did vary expectancies—we can see from records of what the experimenter did or know from the fact that the procedures were automated that expectancies were in fact delivered as we wanted. Yet, subjects did not "catch" that part of the instructions, did not seem to hear, believe, or perceive what we did and said. An additional experiment would be required to test the hypothesis under conditions where the instructions were much more salient and expectancies in fact varied between the groups.

On the other hand, if the subjects had perceived the instructions and the dependent measures reflected no group differences, this would suggest that the intervention was, in fact, manipulated and did not affect the dependent measures outcome. In such a case, the investigator would be more justified in noting that the original hypothesis at least was tested. The adequacy of the test was partially demonstrated by showing that the subjects could distinguish the conditions to which they were assigned. There might well be better tests and stronger manipulations of expectancies that could be provided. But that is another matter.

12.3.2: Keeping Conditions Distinct

Another way in which checking on the manipulation is useful is to ensure that the experimental conditions are, in fact, distinct. The investigator may intend to administer different conditions, instruct experimenters to do so, and provide guidelines and specific protocols of the procedures to ensure that this occurs. Yet the normal processes and interactions of the research assistant and the subject manipulation may override some of the procedural distinctions envisioned by the investigator.

One place where conditions may not remain distinct is the evaluation of different therapy techniques. Part of the problem may be inherent in the subject matter and the way in which it is studied. In therapy investigations, the different techniques often are insufficiently specified and thus the defining conditions and ingredients supposedly responsible for change are not distinguished among groups. Without sufficient specificity, nondistinct global procedures or loosely defined conditions (e.g., "supportive psychotherapy," "mindfulness," or "cognitive behavior therapy") are implemented. There is nothing to criticize about these treatments per se, yet the way they are implemented and perhaps monitored may have blurred them unnecessarily.

Even when the treatments are well specified, blurring and overlap may occur for one of two reasons:

1. The therapists who deliver the treatment may introduce versions or components of one technique while they are administering the other. Perhaps therapist verbal excursions in the sessions or homework assignments of one condition (e.g., mindfulness) are very much like the homework assignments of the other condition (e.g., cognitive behavior therapy) even though that was not intended to happen that way. Now the two treatments are likely to overlap procedurally more than they were designed to at the outset of the project. We discussed the general problem in two different contexts previously in relation to threats to internal validity (diffusion of treatment) and treatment integrity (earlier in this chapter).
2. Two different therapies when administered may genuinely share some common core features (e.g., supportive comments from a therapist, positive social interaction, a relationship or alliance). These are pretty much a common core of many therapies where individuals are seen in person by a therapist.

   Overlap per se may not be detrimental as long as the areas that distinguish treatments are specified and corroborated by a manipulation check.

Were the treatments implemented correctly, and did they remain distinct along the supposedly crucial dimensions specified by their conceptual and procedural guidelines?

> *Treatment differentiation* refers *to showing that treatments in a study of two or more treatments were distinct along predicted dimensions.*
>
> Ensuring that the treatments are distinct (different on key characteristics) is somewhat different from ensuring that the treatments were administered as intended (treatment integrity).

For example, in comparing interpersonal psychotherapy and cognitive-behavior therapy, measures (e.g., ratings of audio or videotapes of selected sessions and coding therapist verbal statements) of how much the therapist focused on interpersonal roles and relationships versus cognitions may show that the treatments were in fact different in what the therapist did. That is, interpersonal therapy sessions may have had significantly more discussion, time, and therapist verbalizations of role-related topics than the cognitive-behavioral treatment, and the reverse pattern may also be evident showing that for time spent on cognitions, the cognitive-behavioral treatment was higher. This is important, but it is still possible that one or both treatments were not administered as intended. It may be that one or more of the therapies suffered a significant departure from the treatment manual, there was a diffusion

of treatment, or sessions were omitted for some of the clients, even though the treatments were distinct.

> In terms of what to remember from this discussion, *treatment integrity* is the key concept. The reason is that this is the usual place that studies break down either in lapses of treatment integrity or not measuring to ensuring that there was integrity.

Treatment differentiation is more specialized and a component of treatment integrity. Therapy studies have reported difficulty in keeping techniques distinct. In classic studies of psychotherapy, behavior therapy, and psychoanalysis that exerted enormous influence, treatment integrity and differentiation were huge problems and could easily explain the no difference findings that dominated the results (e.g., Sloane, Staples, Cristol, Yorkston, & Whipple, 1975; Wallerstein, 1986). Fast forward to now and what is different is the better specification of treatments in manual form to guide therapists more concretely. However, this is only part of the solution. Having a manual does not guarantee adherence to it. Therapists who administer different treatment conditions may include similar elements in both conditions despite efforts to keep treatments distinct. Comparisons of different treatments can be illuminated greatly by gathering information to ensure that the treatments are conducted correctly (integrity) and did not overlap (differentiation) more than might be expected from any common elements associated with therapy or interventions in general.

12.4: Interpretive Problems in Checking the Manipulation

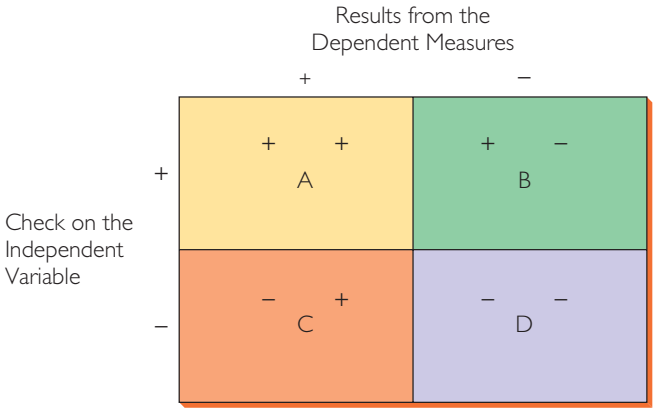
12.4 Analyze four possible situations involving the manipulation check and the dependent measure with respect to interpretive issues

Checking the effects of the manipulation can provide important information that not only aids interpretation of the findings but also may provide important guidelines for further research. The increase in information obtained by checking on the manipulation and its effects has some risk, and this has to be weighed. Discrepancies between what is revealed by the check on the manipulation and the dependent measures may introduce ambiguities into the experiment rather than eliminate them. To convey the interpretive problems that may arise, it is useful to distinguish various simple patterns of results possible in a hypothetical experiment.

Consider a hypothetical experiment that checks whether the independent variable was in fact implemented as intended. After this manipulation check, subjects may complete the dependent measures. When the results are analyzed, it is possible to infer whether the manipulation was implemented effectively from two sources of information, namely, the assessment of the independent variable manipulation check and the dependent measures. These two sources of information may agree (e.g., both suggest that the manipulation had an effect) or disagree (e.g., where one shows that the manipulation had an effect and the other does not). Actually, there are four possible combinations, which are illustrated as different cells in Figure 12.1. For each cell, a different interpretation can be made about the experiment and its effects.

Figure 12.1: Possible Agreement or Disagreement between the Manipulation Check and Dependent Measures

A "+" signifies that the measure shows the effect of the manipulation or that experimental conditions differ on the dependent measures. A "–" signifies that the measure does not show the effect of the manipulation or that experimental conditions do not differ on the dependent measures.



12.4.1: Effects on Manipulation Check and Dependent Measure

This first cell (Cell A) is the easiest to interpret. In this cell, the manipulation had the intended effect on the measure that checked the manipulation (e.g., subjects believed the instructions or performed the tasks as intended or the treatment was delivered as appropriate to the condition). Moreover, the independent variable led to performance differences on the dependent measures (e.g., subjects scores varied as predicted or improved). For present purposes, it is not important to consider whether the predicted relation was obtained but only that the independent variable was shown to have some effect on the dependent variable.

In Cell A, the check on the manipulation is quite useful in showing that the procedures were executed properly but certainly is not essential to the demonstration. The positive results on the dependent measures, particularly if they are

in the predicted direction, attest to the effects of the independent variable. Because of the consistencies of the data for both the manipulation check and dependent measure, no special interpretive problems arise.

12.4.2: No Effect on Manipulation Check and Dependent Measure

In Cell D, there also is little ambiguity in interpreting the results. However, the check on the manipulation greatly enhances interpretation of the investigation. In this cell, the check on the manipulation shows that the independent variable did *not* have the desired impact. The experimental manipulation was somehow missed by the subjects or was too weak to show on the manipulation check. Consequently, the lack of changes on the dependent measures might be expected. The investigator predicted changes on the dependent measures on the presumption that the experimental condition was effectively manipulated.

The pattern of results is instructive because it suggests that additional work is needed to perfect the experimental manipulation (e.g., make it stronger, more salient). The hypothesis of interest really was not tested. That is, the hypothesis was something like, "When I manipulate or change *x* (the independent variable), *y* (the dependent variable) will change." The manipulation check suggested that *x* was not really changed. The results are clarified by showing that the absence of the predicted effects of the independent variable might have resulted from providing a very weak manipulation.

12.4.3: Effect on Manipulation Check but No Effect on the Dependent Measure

Now things begin to become murky. In Cell B, the manipulation check revealed that subjects were influenced by the experimental condition, but the dependent variable did *not* reflect any effect. This is equivalent to the medical cliché that "the operation was a success, but the patient died." This means that the intervention was done well or correctly, but it did not work—not something we as patients are thrilled to hear. (As patients, we want Cells A or C where the outcome is fine no matter how we get there.) The conclusion that would seem to be warranted was that the intervention was well manipulated but that the original hypothesis was not supported. In fact, there may be no relation between the independent variable and the dependent measure, and perhaps this experiment accurately reflected this situation.

> Failure to demonstrate an effect on the dependent measures despite the fact that the manipulation check reveals that the independent variable was successfully implemented does not prove the absence of a relation between the independent and dependent variables.

There are more nuances here than in the Cells A and C we discussed previously. It is possible that the manipulation was strong enough to alter responses on the measure of the manipulation but not strong enough to alter performance on the dependent measures. Some measures may be extremely sensitive to even weak manipulations and others only to very strong manipulations.

For example, in social psychology, we have learned long ago that prejudice is more readily reflected on verbal self-report measures than on measures of overt behavior (e.g., Kutner, Wilkins, & Yarrow, 1952; La Piere, 1934). When individuals are asked whether they will discriminate against or not interact with others, they may readily express such negative intentions. Yet, when these same individuals are placed in a real situation in which they have to exhibit an overt act to discriminate against others, they are much less likely to show prejudicial behavior.

> In other words, prejudice of a given individual or several individuals may vary as a function of how it is assessed.

Alternatively, the strength of prejudice might be defined in part by the extent to which it is evident across different situations and measures. Weak or slight prejudice might be shown only across a few situations and strong prejudice across diverse situations and measures. In relation to Cell B, perhaps it was easy to reflect change on the manipulation check, but not quite enough to show broader and consistent impact on the dependent measures.

The pattern of results in Cell B may indicate that there is no relation between the independent and dependent variable. On the one hand, it may indicate that the manipulation was not sufficiently strong or not implemented in a particularly potent way. If the investigator has reason to believe that the manipulation could be strengthened, it might be worth testing the original hypothesis again. On the other hand, there must eventually be some point at which the investigator is willing to admit that the hypothesis was well tested but not supported. Discussed further below is how to strengthen manipulations to ensure that a strong test of the hypothesis is provided.

12.4.4: No Effect on the Manipulation Check but an Effect on the Dependent Measure

In Cell C, the check on the manipulation suggests that the independent variable was not well manipulated, but the dependent measures do reflect the effect of the manipulation. In this situation, the experiment demonstrated the effects of the independent variable, but ambiguity is

introduced by checking the manipulation. If this were to happen, the investigator would probably regret to have checked the effects of the manipulation at all (and understandably would tear out this chapter of this text).

The task of the investigator is to explain how the manipulation had an effect on the dependent measures but not on the check of the manipulation. The dependent measures, of course, are the more important measures and have priority in terms of scientific importance over the measure that checked the manipulation. Yet the haunting interpretation may be that the dependent measures changed for reasons other than the manipulation of the independent variable. There is no easy way to avoid that interpretation even though it is not the only interpretation or even the most plausible.

One reason that the dependent variable(s) may have reflected change when the available evidence suggests that the independent variable was not manipulated effectively pertains to the nature of statistical analysis.

> It is possible that the differences obtained on the dependent variable were the results of "chance." The results may have been one of the instances in which the subjects' responses between groups were different, even though there is no real relation between the independent and dependent variables in the population of subjects who might be exposed to the conditions of the experiment.

In short, the null hypothesis of the original experiment, i.e., that groups exposed to the different conditions do not differ, may have been rejected incorrectly (a Type I error). The probability of this error occurring in an infinite number of tests is given by the level of significance used for the statistical tests (α).

Another reason that the manipulation check failed to show group differences may be that there were inadequacies with the measure designed to check on the manipulation. The most obvious question that arises is whether the manipulation check measure assesses the construct reflected in the independent variable. Usually manipulation check assessment devices are based upon "face validity," i.e., whether the items seem to reflect the investigator's interest. (*Face validity* has another definition from the one I provided earlier; this is the psychologist's term to justify the basis for using specific items on a measure or a measure itself when in fact no good validation evidence has been obtained. Presumably, the reason this is called "face" validity is to emphasize how difficult it is for us to face our colleagues after having established the validity of an assessment device in such a shoddy fashion, especially when we know better.) There is an uncanny discrepancy in the measures. As investigators we carefully select dependent measures and worry about reliability and validity of the measures we select. Then we compose (other words—slop together, make up) some manipulation check measure that has not type of reliability or validity behind it. Understandably, manipulation check measures are not standardized because studies and the manipulations they use vary widely in research and researchers normally would not be interested in making a mini-career out of developing manipulation check measures.

Other assessment problems with manipulation checks may explain why differences were not found across experimental conditions. For example, the items to assess the manipulation may have been too obscure or unclear. The subjects may have heard the information about the intervention but not have realized its relevance for the assessment device. Alternatively, the variability of the responses to the measure may have been great, leading to the absence of statistically significant group differences. Moreover, the information might not be recalled for manipulation check (e.g., if fill-in items or essay questions were asked) but yet be easily recognized if questions were asked in another way (e.g., multiple-choice questions). Whatever the reason, the failure of the manipulation check to agree with the changes in the dependent measure will interfere with interpretation of most results.

> In any case, one interpretation is that the manipulation check measure was not very good. The measure may have been insufficiently sensitive to pick up the manipulation and perhaps was not even a good measure of the manipulation check (low reliabilities and validities).

It is quite possible for the manipulation check to reflect some other construct than the independent variable. It is of little consolation to raise this as a possibility after an investigation is completed. The manipulation check is part of the methodology for which the investigator can rightly be held responsible. Hence, prior to the experiment, it is important for the investigator to have some assurance that the manipulation check will reflect actual differences across conditions. Reflecting change on the measure can be accomplished in pilot work prior to the full experiment as a minimal validation criterion of the assessment device.

12.4.5: General Comments

Pointing out the ambiguities that can result from checking how successful the independent variable was manipulated could discourage use of such checking devices. This would be unfortunate because much can be gained from knowing how effectively the independent variable was manipulated. Such checks, as a supplement to information on the dependent measures, provide feedback about how well the hypothesis was tested.

> A failure to achieve statistically significant group differences on the dependent measures is instructive but does not convey specific details about the experimental manipulation.

Changes in dependent measures reflect many events all working together, including whether the manipulation was implemented effectively or was potent enough, whether the measures were appropriate for the manipulation, and whether procedural errors were sufficiently small to minimize variability. The absence of effects on dependent measures could be attributed to many factors, only one of which is the failure to implement the independent variable effectively. On the other hand, a manipulation check helps provide more specific information and hence can be very useful in interpreting a given study and guiding subsequent studies.

12.5: Special Issues and Considerations in Manipulation Checks

12.5 Analyze some of the main point areas that arise due to ambiguity on how and when to use manipulation checks

Many issues emerge in deciding how and when to use manipulation checks.

12.5.1: Assessment Issues

One assessment issue relevant for deciding whether to check on the manipulation is the possible reactivity of assessment and the relevance of reactivity for the particular experiment. By checking on the manipulation, an experimenter may arouse subjects' suspicions about the experiment and raise questions that ordinarily might not arise. The manipulation check may even sensitize subjects to the manipulation.

For example, a self-report questionnaire to check on the manipulation may make the manipulation that just occurred more salient to the subject. As an extreme case, the experimental manipulation might consist of altering the content of a subject's conversation during a standard interview as a function of events that happen to the subject in the waiting room prior to the interview.

> *Confederates*, persons who work for the investigator, may pose as other subjects innocently waiting their turn in the waiting room but, in fact, engage in prearranged discussions designed to influence the subject.

The prearranged discussions would vary across subjects depending upon the exact experimental conditions. To check on this manipulation, the investigator could ask subjects at the beginning of the interview such questions as what they talked about in the waiting room or what their current mood is. The questions might suggest to the subjects that their previous interaction in the waiting room was part of the experiment and arouse suspicions and reactions that would not otherwise be evident if no manipulation check were used.

As a general point, reactivity of the manipulation check per se might not be important depending upon how the investigator conceives the manipulation and the process through which it affects the subject. Yet in some circumstances, the experimenter might not want to risk the likelihood that the manipulation check itself changes the subject in some way. If the check is important, the investigator may wish to design unobtrusive measures that are less likely to arouse suspicions than are direct self-report measures. For example, the experimenter might leave the subjects alone with another subject (actually a confederate) who asks, "Say, what is this experiment about anyway?" or "What did the experimenter say about what's going to happen?" Responses to a few such questions could be scored (e.g., from audio tapes, through a one-way mirror, or by the confederate) to address the question whether the subjects perceived the purpose of the study or to assess other specific aspects of the manipulation.

Alternatively, the investigator may administer the manipulation check after the dependent measures are assessed. Even if the manipulation check is reactive, this could not influence the results because the dependent measures have already been completed. The disadvantage with this alternative is that the longer the delay between the manipulation and assessment of the manipulation's impact, the greater the chances that the check will not discriminate groups. During the delay, subjects may forget precisely what they heard in the instructions or original rationale. Also, it may be possible that the dependent measures, if completed first, could influence the results on the manipulation check.

12.5.2: More Information on Assessment Issues

> To avoid reactivity of the manipulation check, the investigator might simply assess the manipulation and its effects in pilot work prior to the investigation.

In pilot work, self-report questionnaires to assess the manipulation can be used without even administering the dependent measures. In addition, the investigator will have a good basis for knowing in advance that the independent variable was effectively manipulated.

The decision whether to check on the effects of the manipulation also pertains to whether subject awareness of the independent variable is at all relevant. We already know outside of methodology that many influences on us in everyday life are well below our conscious awareness. For example, sexual attraction is influenced by scents (e.g., that reveal testosterone), shades of skin color, and color of clothing worn, among many other such factors (e.g.,

Beall & Tracy, 2013; Thornhill, Chapman, & Gangestad, 2013). For example, women are more likely to wear red or pink when at peak fertility, but this is not a clear guide and not something men who are attracted to women can identify (e.g., not all or even most women wear those colors when at peak fertility and some wear those colors when not at peak fertility). This area, riddled with interesting research, conveys multiple influences that are nuanced, change over time, and probably have strong cultural determinants as well. **But the key**—we are rarely aware of what the influences are.

Back to methodology. Effective experimental manipulations do not necessarily operate through subject's awareness. For some manipulations, it may be entirely irrelevant whether subjects know or could recognize what has happened to them in the experiment. In social psychology, a great deal of research focuses on the topic of priming conveys how important experimental manipulations promote (prime) behavior well out of the awareness of the subjects. In this work, cues in the environment are experimentally manipulated (e.g., holding a warm coffee cup, smelling cleaning liquids, seeing a person who is elderly walking with effort) (Bargh, 2007). These cues are placed in the environment in a way that subjects to not see them as part of an experiment (e.g., are momentarily asked to hold a cup of warm coffee while the experimenter is juggling other materials in his or her hands). They do not consciously perceive the cues or connect them to the experiment or to their own behavior. Actually, they perceive them very well but at a level below consciousness. When a manipulation check or equivalent is administered, such as asking subjects to explain why they did this or that, they do not recognize the intervention cues. The participants readily were influenced by the cue as reflected on the dependent measures of the study but unbeknownst to them (do not know what the manipulation was consciously).

> Here the manipulation check does not show the effect but the dependent measures do and that is the purpose of the research, i.e., to show behavior is readily altered by cues that are not recognized at all (on a manipulation check).

Priming is mentioned to make a more general point. The fact that participants do not show awareness on the manipulation check measure does not necessarily mean the manipulation did not register or have impact. As one ponders the use of manipulation checks in a given study, it is important to ask, "If someone does not indicate the manipulation registered on some manipulation check measure, does that necessarily mean the independent variable had no impact?"

What do you think?

It is rare that one could answer that yes. This is not trivial in part because investigators routinely throw out subjects from a study because they did not score predictably on the manipulation check, a point we discuss next. More generally, the type of manipulation determines the manner and focus of the manipulation check. This means that in some cases subject perceptions are relevant and in others of ancillary importance.

12.5.3: Data Analysis Issues: Omitting Subjects

The discussion has presented the notion that an intervention is or is not effectively manipulated as determined by a check on the manipulation. It is unlikely that effectively manipulating an independent variable is an all-or-none matter. The manipulation will usually not succeed or fail completely but will probably affect subjects within a given condition differently. A given proportion of subjects may be affected by the manipulation. This proportion could be defined by answers to particular questions. For example, subjects who answer most (e.g., >80%) or all questions about the experimental manipulation correctly may be considered those for whom the condition was successfully implemented. Whatever the criteria, usually there will be some people for whom the experimental condition was effectively manipulated and others for whom it was not, as operationalized by the manipulation check measure. An important methodological and practical question is how to treat subjects in the data analyses who are differentially affected by the manipulation. Consider the options.

> *Delete the Participants from the Data*. At first blush, it seems reasonable to include in the analyses only those subjects who were truly affected by the manipulation. After all, only those subjects provide a "real" test of the hypothesis.

That is, my hypothesis was subjects who received my very clever manipulation would no longer be depressed and would join a monastery. Why would I want subjects in this study who for whatever reason were not paying attention, were slow mentally, and just did not get it, and so on. Of course, I should dump them.

But wait. Random assignment to groups was how the study began and that is not trivial or merely a methodological nicety to impress others. Random assignment was likely to disperse diverse confounding variables (we called them "nuisance variables") across conditions in a nonsystematic way. That is great. Merely using subjects who show the effects of the manipulation on the manipulation check measure may lead to select groups of subjects that vary on several characteristics from the original groups that were formed through random assignment. Deleting subjects from the study violates the randomness of the assignment procedure and could lead to selection bias, a threat to internal validity. At the end of the study, you might say that the

manipulation was the basis of group differences. I would say selection bias is a threat to internal validity and cannot be ruled out or made implausible. We have spoken about attrition or loss of subjects in an experiment as a basis for selection bias. In this case, deleting subjects is caused by the investigator who omits subjects from the analyses based on the manipulation check measure. Be wary when you do research or read research where subjects are just tossed. There might be fabulous reasons, so just at this point in our discussion, wariness is fine.

Bear in mind other considerations as well:

1. The manipulation check measure is not usually a well-developed measure and is hard to defend as a reliable or valid assessment device. I would be very hesitant to throw away the benefits of randomness based on a home-made unreliable measure with only face validity!
2. Omitting subjects obviously means reducing the total number in the study.

   Most studies in psychology do not have an adequate sample size to detect differences (are underpowered), and the last thing one wants to do is to throw out some more and make the sample smaller.

3. Failing to do "well" on the manipulation check does not mean the manipulation was ineffective. The primary index of whether the manipulation was effective is the set of dependent measures. Deleting subjects based on the manipulation check does not mean one is deleting subjects who failed to respond on those measures. The assumption that change on the manipulation check is a precondition for change on the dependent measure seems to follow common sense, but actually has no strong basis.

12.5.4: More Information on Omitting Subjects

In clinical psychology and education, there is a special case where participants occasionally are deleted because they did not receive the manipulation. For treatment, prevention, and educational studies, the "manipulation" is a regimen of some intervention well beyond one quick session. For example, intervention studies often are conducted in the schools to prevent problems such as:

- Bullying
- Child maladjustment
- Cigarette smoking
- Drug abuse
- Teen pregnancy
- Suicide

The schools represent an opportune setting for such interventions because students are available for extended periods (e.g., months), school attendance is required, teachers can integrate interventions in the classroom, and interventions can be administered on a large scale (e.g., several classes or schools). In large-scale applications, treatment integrity is difficult to achieve and sustain. Consequently, at the end of prevention trials in the schools, large differences can be evident in the fidelity with which classroom teachers implement the interventions. Invariably, some teachers carry out the procedures extremely well, others less well, and still others not at all. "Carrying out the procedures well" is equivalent to a manipulation check in the sense that for the classes of these teachers, the manipulation was delivered well (as measured in some objective way). At the end of such a study, investigators occasionally exclude classrooms (teachers and subjects) where the intervention was not carried out or carried out well. Again, it may seem reasonable to exclude teachers (and their classes) in the intervention group. After all, these teachers did not conduct the intervention or did not meet minimal criteria for delivery of the intervention. The investigator is interested in evaluating the effect of the intervention when implemented or implemented well relative to no intervention. Thus, the investigator selects only those intervention classes where the program was well delivered. These teachers and classes are compared to nonintervention classrooms that, of course, did not receive the program.

Unfortunately, selecting a subgroup of teachers who carried out the intervention with fidelity violates the original random composition of intervention and nonintervention groups or conditions in the study. Data analyses of the selected intervention group and nonintervention group now raise threats to internal validity (namely, selection *x* history, selection *x* maturation, and other selection variables that apply to one of the groups) as a plausible explanation of the results. Group differences might simply be due to the special subset of teachers who were retained in the intervention group—these teachers are special. Alternatively, it may be that the teachers who adhered better to the intervention had classes and students who were more amenable to change or in someway more cooperative. That is, adherence may have been easier because of the students in the specific classes. In short, when one omits classes and subjects, it may not be the integrity of the intervention that is being evaluated as much as it is the specialness of the teachers or students who adhered to the procedures.

12.5.5: Intent-to-Treat Analyses and Omitting and Keeping Subjects in Separate Data Analyses

The most appropriate analysis of results is to include all subjects who were run in the various experimental conditions ignoring the fact that only some of them may

have shown the effect of the intervention on the manipulation check.

> An analysis that includes all subjects provides a more conservative test of the intervention effects. The inclusion of all subjects in this way is often referred to as *intent-to-treat analyses*.

Intent to treat usually is used in the context of subjects who drop out of a study (e.g., before posttreatment or follow-up assessment). One includes subjects in the data analyses and uses the last data that the subjects have provided (e.g., pretreatment data) for any subsequent analyses (e.g., posttreatment). Thus, if subjects dropped out of treatment, the pretreatment data would be used as both pre- and posttreatment scores. Intent-to-treat analysis has two virtues. First, it allows inclusion of all subjects even those who did not complete the study. This preserves the random assignment and hence makes any selection biases unlikely. Second, the analyses provide a conservative test of the hypotheses. The analysis has some cautions too. I mention the situation here because sometimes subjects are omitted for manipulation check sorts of reasons related to not receiving the intended intervention in the intended fashion.

> *Ok, I am Persuaded. Keep the Participants in the Study.* Ok, you succumb to the reasoning and you keep all subjects in study. Now you have retained the random composition of groups and are appropriately proud of yourself. You analyze the results and find no differences between groups.

At this point, few humans can resist the temptation to go back and reanalyze by tossing subjects who did not respond to the manipulation check. That brings all of the problems we mentioned previously. There is an alternative within this option.

Test your view. That view is that people who responded to the manipulation check did better, showed higher scores, or whatever on the dependent variables. This is an empirical question. Using all subjects, compute a correlation between the scores on the manipulation check and scores on the dependent measures. (If there was a pretest and posttest, look at scores on the manipulation check and amount of change.) Analyses like these use all of the data and address in a quantitative way (correlation coefficient for example), the relation between scores on the manipulation check and dependent measures. If that relation is high, you have something to talk about. If that relation is low, it is good that you did not delete subjects. It is not clear what the manipulation check measured, but it had little relation to the dependent measures.

> **Do Both***: Omit and Keep the Subjects in Separate Data Analyses*. It is important to include and report all of the data analyses in any report of the study and to preserve randomness whenever possible. Run the data analyses in two ways. In the first run, evaluate all of the subjects included in the study whether or not the manipulation check shows they experienced, understood, etc., the manipulation. In the second run, now exclude those subjects who you believe were flaky on the manipulation check measure.

Specify the criterion for that, justify why you chose that criterion, so it is not based on a thousand analyses and looking for the one that came out to be statistically significant. In a kind world, you may find that both analyses yield the same results. If there are differences, you can speculate why, suggest further research, and so on. If only the manipulation check responders changed on the dependent measures, you may be able to do some further analyses to suggest that they were or were not different on other measures (e.g., sex, IQ, socioeconomic status). There may be a selection variable that you can identify and even control (statistically).

There is one more type of analysis that may be informative and guide research—your own research. Take all of the subjects in the experimental manipulation group and place them into one of four groups. Please go back to Figure 12.1 to see what those groups are. Essentially, we want to identify those who responded well to the manipulation check and also showed the predicted effect on the dependent measures (Cell A in Figure 12.1) but all the other possibilities as well (Cells B, C, D). Once we form those groups, now analyze any subject and demographic variables that you collected. Can we identify other variables that sort out who responded and who did not (2 x 2 analyses of the four cells)? This is a post-hoc analysis and one ought to be careful, but it is also a search for moderators that might help in planning the next study.

12.5.6: Pilot Work and Establishing Potent Manipulations

Establishing the efficacy of an experimental manipulation probably is best accomplished in preliminary (or pilot) work before an investigation, especially if an investigator is embarking in an area of research in which he or she has not had direct experience.

> *Pilot work is a test or preliminary effort to evaluate aspects of the procedures to see if they work (e.g., equipment, recruitment methods), are feasible, and are having the effect (e.g., on the manipulation check or even dependent measures) before the full study is run.*

The actual study itself should be based on preliminary information that the manipulation can be implemented effectively. Preliminary or pilot work to learn how to manipulate the independent variable successfully can be invaluable for the subsequent results of a research program.

Pilot work usually consists of exploring the intended manipulations by running a set of subjects who may or may not receive all the conditions and measures that will be used in the subsequent experiment itself.

In pilot work, the subjects can contribute directly to the investigator's conception and implementation of the manipulation. For example, the subject can receive the manipulation and complete the manipulation check. At this point, the subject can be fully informed about the purpose of the investigation and provide recommendations about aspects of the procedure that appeared to the subject to facilitate or detract from the investigator's overall objective. Detailed questions can be asked, and the ensuing discussion can reveal ways in which the manipulations can be bolstered. An increasingly common practice is the use of focus groups, i.e., meetings with groups of individuals who are knowledgeable in light of their special role (e.g., consumers, parents, teaches, adolescents) to identify what is likely to have impact in a particular area.

> Focus groups are away to obtain opinions, often informally, about a question of interest. Meeting with groups of individuals before designing a manipulation (e.g., remedial program) or after running a program can be useful for generating concrete ideas to improve the intervention as well as ways to assess the manipulation.

Another reason for developing the manipulation in pilot work is that some of the problems of checking the manipulation can be eliminated. As discussed earlier, checking on the manipulation in an experiment may sensitize subjects to the manipulation and presumably influence performance on the dependent measures. Pilot work can check the success of the manipulation with, for example, self-report questionnaires. There is no need to obtain measures of performance on the dependent variable if preliminary work is to be used merely to establish that the experimental conditions are administered effectively. If the manipulation has been shown to be effectively manipulated in pilot work, the investigator may wish to omit the manipulation check in the experiment to avoid the possibility of sensitization effects. Of course, a pilot demonstration does not guarantee that the experiment will achieve the same success in manipulating the independent variable, because subjects differ in each application. However, a pilot demonstration can greatly increase the confidence that one has about the adequacy of the experimental test.

Pilot work can be very useful in advance of an investigation to develop distinct experimental conditions consistent with the desired manipulation. An investigation usually reflects a considerable amount of effort and resources, so it is important to ensure that the independent variable is successfully manipulated and the experimental conditions are as distinct as possible. If the manipulation is weakly implemented and doubt can be cast on the effectiveness with which the hypothesis is being tested, the interpretation of the final results may leave much to be desired.

12.6: Assessing Clinical Significance or Practical Importance of the Changes

12.6 Examine how the interventions are evaluated with respect to treatment, prevention, education, and enrichment programs

A major area of assessment in clinical psychology and other mental health professions is the evaluation of interventions, as reflected in treatment, prevention, education, and enrichment programs. As in any study, intervention research has a set of measures to evaluate the impact of the "manipulation," i.e., intervention to evaluate change. Thus, treatment studies of anxiety, eating disorders, and drug use, for examples, may assess measures of the target problem and perhaps as well ancillary effects such as other domains of improvement (e.g., multiple symptom areas). Yet, intervention research raises special assessment beyond those raised in evaluating experimental manipulations in a laboratory study.

Intervention research has scientific goals (e.g., addressing a question that remains to be resolved; evaluating some nuance of treatment or comparing treatments) but also applied goals of being of direct benefit to clients or patients who participate in treatment. The scientific question is addressed by showing changes, differences, and effects on the dependent measures whatever they are. The applied question may be reflected on those same measures, but that is the part that may be questionable.

> Do we know that our interventions are making a difference in the lives of the individuals?

Treatment research evaluates the effects of interventions by showing statistically significant changes from pre- to post-treatment (e.g., reduction in symptoms of depression) and statistically significant differences (e.g., one treatment is better than another). The strength of the effect also may be reported as an effect size where the magnitude of the difference is quantified. Statistical significance and effect size do not address the question of the applied importance of the outcome or effect. We want to know whether the intervention makes a difference that has impact on the client. Clinical significance is designed to address this question.

> *Clinical significance refers to the practical value or importance of the effect of an intervention, i.e., whether it makes any "real" difference to the clients or to others in their functioning and everyday life.*[1]

For example, consider that we conduct a study with young adults referred for severe anxiety (e.g., fear of open spaces or debilitating anxiety in social situations). We have a proper control group (e.g., treatment as usual) to address threats to internal and construct validity, recruit a large number of participants to handle power, evaluate treatment integrity—we do it all! In the end, we find those treated with the special condition improved on our outcome measures (statistically significant change) and when treatment and control conditions are compared the effect size is large (Cohen's d = .8). Now we want to ask the question of clinical significance. Outside of the sessions and beyond our reliable and valid measures, has treatment had any impact on measures of everyday functioning? For example, can the participants go out of the house, participate in social situations, or enjoy life in ways that were somehow precluded by their anxiety?

Improvement on psychological measures including the usual types of measures (objective measures, self-report, biological indices) may not be sufficient to establish that. It is possible to change on various inventories and questionnaires without any of these other benefits. The reason is that most measures have all sorts of reliability and validity but that alone does not translate to practical changes that make a difference. Thus, showing that depression scores on some scales improved is likely to be related to functioning in everyday life, but we do not know exactly how much or who that translates to actual affect, cognition, behavior, socialization, work functioning, and so on in one's life. Consequently, there is a deep concern about how to measure treatment outcome and as a part of that what indices might convey genuine impact, and this is an ongoing concern (e.g., De Los Reyes, Kundey, & Wang, 2011; Hill et al., 2013; Ogles, 2013; Verdonk et al., 2012). Clinical significance raises the concern about changes in these other areas (e.g., in daily life) but also about changes in the target area (anxiety). Many different indices to operationalize clinical significance have been used or proposed. Table 12.1 summarizes strategies in use and relevant to evaluation of

Table 12.1: Measures to Evaluate the Clinical Significance of Change in Intervention Studies

Method	Defined
Falling within Normative Levels of Functioning	At the end of treatment, evidence on the measure that clients now fall within the range of a nonclinical group on a measure. The usual measure is one of the outcome measures used in the study to reflect symptom change. The added information for clinical significance is data obtained from a normative sample on the same measure. A range (e.g., one standard above and below the mean) is selected to define a normative level. This can be an arbitrary range or derived statistically but scores that statistically provide the best estimate of separating clinic and nonclinic samples. Evidence for clinical significance is that clients were outside that range before treatment but fall within that normative range after treatment.
Magnitude of Change the Clients Make from Pre- to Posttreatment (Reliability Change Index)	A large improvement from pre- to posttreatment on the outcome measures. A commonly used criterion is a change from the mean of the pretreatment scores. Thus, at posttreatment individuals whose scores depart at least 1.96 standard deviations from the mean of their pretreatment scores would be regarded as having changed in an important way.
No Longer Meeting Diagnostic Criteria	In many treatment studies, individuals are recruited and screened on the basis of whether they meet criteria for a psychiatric diagnosis (e.g., major depression, posttraumatic stress disorder). A measure of clinical significance is to determine whether the individual, at the end of treatment, continues to meet criteria for the original (or other) diagnoses. Presumably, if treatment has achieved a sufficient change, the individual would no longer meet criteria for the diagnosis.
Subjective Evaluation	Impressions of the client or those who interact with the client that indicate that changes make a palpable difference in the function of the client difference. Ratings of current functioning and whether the original problem continues to be evident or affect functioning.
Clinical Problem Is No Longer Present	At the beginning of treatment, there was a clear problem (e.g., panic attack, tic, uncontrolled nightmares); at the end of treatment, the problem is gone completely. This is a quantitative change (e.g., high score on a symptom measure to zero) but better conveyed as a qualitative change. The person no longer has an episode or the frequency is so low (e.g., once every 6 months rather than 6 times a day) that the effect is stark and obvious.
Recovery	Functioning well and managing many different spheres, including health (living in a healthy way or managing one's condition), home (having a stable and safe place to live), purpose (e.g., having a purpose and being involved in meaningful activities), and community (e.g., having relationships with others and social networks with support, friendships, love, and hope).
Quality of Life	Quality of life refers to the clients' evaluation of how they are doing in several spheres of life (e.g., health, philosophy of life, standard of living, work, recreation, love relationship, relationships with one's children, friendships, and others); overlaps with recovery in foci but the orientation is not overcoming a disease or clinical problem per se.
Qualitative Assessment	In-depth evaluation of an individual with open-ended questions that better allow the individual to evaluate the impact of treatment and along what dimensions. The assessment can consider how therapy did and did not help and whether any changes make a difference and in precisely what ways for that individual.
Social Impact	Change on a measure that is recognized or considered to be critically important in everyday life; usually not a psychological scale or measure devised for the purposes of research. Change reflected on such measures as arrest, truancy, hospitalization, disease, and death that are of broad interest to society at large in addition to the individual.

the importance of the change. They are highlighted here to clarify the issue, to provide options to include into interventions studies, and to prompt creative thinking about novel measures that accomplish the same goal.

12.6.1: Most Frequently Used Measures

Most treatment outcome studies do not assess clinical significance, and that is an important point so that the heading of this section does not mislead. That said, when clinical significance is assessed, one of the following four methods listed in Table 12.1 tends to be used.

12.6.2: Further Considerations Regarding Most Frequently Used Measures

Falling within Normative Levels of Functioning. The question addressed by this method is, "To what extent do clients after completing treatment (or some other intervention) fall within the normative range of performance?"

Prior to treatment, the clients presumably would depart considerably from their well-functioning peers (e.g., a community sample not referred for treatment) on the measures and in the domain that led to their selection (e.g., anxiety, depression, social withdrawal). It is likely that the clients were actually screened in the clinical trial to be sure that they met various criteria for clinical dysfunction (e.g., psychiatric diagnosis of major depression, extreme scores on one or more well-validated measure). Demonstrating that after treatment, these same persons were indistinguishable from or within the range of a normative, well-functioning sample on the measures of interest would be a reasonable definition of a clinically important change.

To invoke this criterion, a comparison is made between treated clients and peers (e.g., same age, sex, perhaps same cultural, ethnic and socioeconomic composition) who are functioning well or without significant problems in everyday life. This requires that the measures used in the study have normative data available from community (nonpatient) samples. Data from the normative sample usually are used to provide a mean and standard deviation. A range (e.g., one standard deviation above and below the mean) is identified and called the normative range. We know from the "normal distribution" that one standard deviation above and below the mean would include approximately 68% of people in the sample; we could make that two standard deviations above and below the mean and that would include 95% of the normative population. This is usually arbitrary, but occasionally research is conducted to identify the point (scores) that seems to predict status as in a clinical versus nonclinic sample. However, the range is formed, and one examines the percentages of individuals who went from out of that range before treatment to within that range after. Yes, statistical regression is a problem. Because clients were selected because of their extreme scores, it is possible that many will show some improvement because of regression. Yet, this is also true of any control group, so regression does not explain any differences in improvement between groups.

A difficulty with this method is the requirement of data about a normative population. If standardized assessments are used for which there is a large database of normative samples (e.g., Beck Depression Inventory, Child Behavior Checklist), then one can draw on that. Yet, in many studies there is no normative database on which investigators can draw, so some other index is likely to be used.

Magnitude of Change the Clients Make from Pre- to Post-treatment. A more common method of defining clinical significance is to look at the magnitude of the changes of the clients without any comparison to a normative group. The criterion here is how much change the individual makes and to set a criterion to determine whether the change is or is not clinically significant. This method is referred to as the *reliable change index* and is calculated separately for each individual. The individual's post-treatment score on a measure is subtracted from the pre-treatment score. The goal of this subtraction is to see how much improvement there is, so which score is subtracted from which is based on the direction of scoring (is higher score at the end of treatment an indication of improving or becoming worse). This difference is divided by the standard error term based on the sample in the study.

The formula for this is:

$$\text{Reliable Change Index} = \frac{A_{post} - A_{pre}}{S_{diff}}$$

Where A = assessment on a particular measure at pre or post and S = the standard error of difference scores

A change of greater than 1.96 is considered to be a clinically significant improvement. This number (1.96) is in standard deviation units and is exactly the criterion used when a statistical test (*t* test) compares groups. The level for a statistically significant effect is a *t* of 1.96, $p < .05$. That is the reason for taking 1.96 and applying to change in the individual's score.

Let us put all of this into words. There is a comparison. At the beginning of the study, all clients screened for dysfunction are considered to define a "dysfunctional sample." At the end of treatment, if a clinically important change is made, scores of a client ought to depart markedly from the original scores of the sample. The departure of course ought to be in the direction of improvement (e.g.,

lower symptoms). There is no logical justification for deciding how much of a change or reduction in symptoms is needed and different criteria have been suggested and used (e.g., Jacobson, Roberts, Berns, & McGlinchey, 1999; Ogles, 2013). In the version I noted previously, a departure that is 1.96 deviations utilizes a criterion already employed in statistical test of significance. This measure of clinical significance asks, "Did the posttreatment scores really depart from the pretreatment distribution of scores (that defines a dysfunctional sample)?" Yet, the focus is on the individual client and her or his improvement rather than performance of the group.

As an example, in one study depressed or anxious adults were randomly assigned to one of two treatments—cognitive behavior therapy (CBT) or acceptance and commitment therapy (ACT) (Forman et al., 2012). At the end of therapy, treatments were no different in their effects. Yet, at follow-up 1½ years later, the effectiveness of CBT was significantly better. One index used to reflect this was the reliable change index. On a standard measure of depression (Beck Depression Inventory), the percentage of clients whose change to meet this criterion was 81.8 and 60.7 for the CBT and ACT treatments, respectively. Other measures also favored CBT using the reliable change index were consistent with this difference as well. These results show use of the measure but also convey an important point about treatment, namely, that conclusions reached about the effectiveness of treatment evaluated at one point in time (e.g., immediately after treatment) may be different from those warranted at a different point in time (e.g., follow-up).

For many measures used to evaluate treatment or other interventions, normative data that could serve as a criterion for evaluating clinical significance are not available. That is, we cannot tell whether at the end of treatment cases fall within a normative range. However, one can still evaluate how much change the individual made and whether that change is so large as to reflect a score that is quite different from the mean of a dysfunctional level (pretreatment) or sample (no-treatment group). Of course, if normative data are available, one can evaluate the clinical significance of change by assessing whether the client's behavior returns to normative levels and also departs from dysfunctional levels.

12.6.3: More Information on Most Frequently Used Measures

No Longer Meeting Criteria for a Psychiatric Disorder. Clinical significance also is assessed by evaluating whether the diagnostic status of the individual has changed with treatment. In many treatment studies, individuals are recruited and screened on the basis of whether they meet criteria for a psychiatric diagnosis (e.g., major depression, posttraumatic stress disorder [PTSD]). Those with a diagnosis are included in the study and assigned to various treatment and control conditions. A measure of clinical significance is to determine whether at the end of treatment individuals continue to meet criteria for the original (or other) diagnoses. Presumably, if treatment has achieved a sufficient change, the individual no longer meets criteria for the diagnosis.

For example, an exposure-based treatment was compared to a wait-list control condition for the treatment of veterans who met diagnostic criteria for PTSD (Church et al., 2013). Apart from measures of symptom change, clinical significance was measured by seeing if there were group differences in diagnosis when treatment was over. At the end of treatment, 90% of treated veterans no longer met criteria for the diagnosis of PTSD compared to 4% of the control conditions. The results are impressive and go beyond showing symptom improvement. The status of individuals (no longer meeting diagnostic criteria) was clearly changed by the intervention program. Statistical regression might lead to some symptom reduction, but this was likely to be equally evident in the control condition, given that veterans were assigned to conditions randomly.

There is something appealing about showing that after treatment an individual no longer meets diagnostic criteria for the disorder that was treated. It suggests that the condition (problem, disorder) is gone or "cured." For many physical conditions and diseases (e.g., strep throat, "the flu," rabies), no longer meeting the diagnostic criteria can mean that the disorder is completely gone. That is, the presence or absence of these disorders is categorical, and "cured" may make sense as a concept. In psychology and psychiatry, we usually say, "only hams are cured." The reason is that we know that many disorders are not all or none. Most disorders are now considered "spectrum" disorders; that is, they are on a continuum and there is not defensible cut point to say one does or does not "have it." Autism, once defined categorically, is now considered autism spectrum disorder as an example most familiar in the news. But the point applies to many other disorders as well. Depression and conduct disorder among many other examples are not present or absent as a cut point, although for many reasons (e.g., insurance reimbursement, access to other benefits) a point may be defined. Thus, there is no magic to meeting criteria or a disorder or just missing it.

In relation to clinical significance what this means is that a change in one or two symptoms (e.g., in degree or present or absent) could lead one to say, "this person no longer meets criteria for the diagnosis" but the person could still be suffering, have problems, and not be doing all that great in everyday life. Individuals who do not quite meet the criteria for the diagnosis but are close can still have current and enduring social, emotional, and behavioral problems (e.g., Touchette et al., 2011; Van Os, Linscott, Myin-Germeys, Delespaul, & Krabbendam, 2009). Consequently, no longer

meeting diagnostic criteria does not necessarily mean the person is appreciably better or that the change made an important difference in the client's life.

> *Subjective Evaluation.* The subjective evaluation criterion for clinical significance usually consists of clients reflecting on their treatment and the changes they have made. Usually a self-report questionnaire is used, and this may be composed of face valid items.

The subjective evaluation method has two components that are worth distinguishing. Both involve someone's subjective judgment, but they differ in foci.

The first of these is used in treatment research already. *This is referred to as subjective evaluation and consists of determining the importance of behavior change in the client by assessing the opinions of individuals who are likely to have contact with the client or in a position of expertise* (Wolf, 1978). The question addressed by this method of evaluation is whether changes have led to differences in how clients and other people see the change. The views of others may be especially relevant in some circumstances (e.g., with children, the elderly). They are relevant because people in everyday life (parents, teachers, adult children) often have a critical role in identifying, defining, and responding to persons they regard as dysfunctional or deviant. Subjective evaluations permit assessment of the extent to which the effects of an intervention are apparent to the clients or to others.

There is a way in which this is obviously important—at the end of treatment, do the clients view themselves as better off? No matter what objective changes were achieved (e.g., less anxiety, more socialization, virtual elimination of self-injurious behavior), clients' view of the value of what they received and how the change affected their lives is very important. We would not only want symptoms and disorders to improve in our clients, but also for clients to experience and see those changes as important.

For example, insomnia and its treatment is an active area of research. Insomnia has a high prevalence in the general population (e.g., 33%) and is associated with both medical and psychiatric conditions (e.g., chronic pain, depression). Outside of these conditions, insomnia is also associated with impaired concentration and memory, irritability, difficulty in interpersonal relationships, decreased quality of life, and increased risk of new-onset psychiatric disorder (see Mitchell, Gehrman, Perlis, & Umscheid, 2012). (Medication is the most common treatment, so there is also concern about use and abuse of that.) Intervention research often uses a variety of objective measures, assessed in a sleep lab where clients are monitored overnight. All sorts of measures are used to evaluate duration and quality of sleep (e.g., sleep latency, wake after sleep onset, sleep efficiency [ratio of time spent asleep/time spent in bed], total sleep time, and total wake time). In the majority of studies, patients are not asked (subjective evaluation)—how is your sleep? Do you feel any better? Did treatment make a difference? In addition to objective measures of sleep, it would be great to just ask participants if they are doing better in their lives after the treatment, especially if there were a validated measure available.

> Subjective evaluation is obviously important. If treatment is working and has an important impact, the effects ought to make a perceptible difference to clients themselves and to those with whom they interact.

The opinions of others in contact with the client are important as a criterion in their own right because they often serve as a basis for seeking treatment in the first place and also reflect the evaluations the client will encounter after leaving treatment. Subjective evaluation is relevant as a criterion.

The second way in which subjective evaluation is important is drawn from a context outside of treatment evaluation. A huge body of research focuses on subjective well-being and happiness (e.g., Diener & Chan, 2011; Lucas & Diener, 2009).

> *Happiness is defined by whether people believe and experience their lives as desirable, rewarding, and satisfying.*

Happiness is related to quality of social relationships, economic prosperity, and health. Interestingly, happiness results from other experiences but also has been shown to cause (lead to) how people relate to others. Happy people become friendlier, more cooperative, and more likely to be productive at work. Thus, happiness is associated with health, social benefits, and overall quality of life. In relation to clinical significance, it would be quite valuable to measure happiness. Among the reasons is that the construct connects to a large empirical literature, reliable measures are available, and we know that happiness relates to multiple domains of functioning. If therapy could improve happiness, that would be quite important. Among the vast array of evidence-based treatments, we do not know how or whether they improve happiness.

12.6.4: Other Criteria Briefly Noted

The measures covered to this point are by far those that are the most frequently used. Yet there are many other methods that are viable options and no solid data to prefer those already noted from the ones highlighted here (and included in Table 12.1). I mention them here briefly because they are not used regularly in randomized controlled trials of psychosocial treatments. Yet each reflects important information about impact that goes beyond the usual measure used in research.

Clinical Problem Is No Longer Present. Sometimes, a clinically significant change can be inferred when a problem is reduced to zero or no longer occurs at the end of treatment.

For examples, at the beginning of treatment the clients may show high rates of binge eating, panic attacks, episodes of self-injury, tics, illicit substance use, and domestic violence. Reducing these to zero by the end of treatment would clearly be an important change. It might be that cutting the rate in one-half for the client is important too, but that can be ambiguous (e.g., reducing gambling, cigarette smoking, or alcohol consumption from some large amount to half of that amount). Clearly for many clinical foci, elimination of the problem is less ambiguous and arguably can be stated to reflect an important change. This criterion bears some resemblance to no longer meeting diagnostic criteria, but the specific diagnostic criteria (cutoff point that meets the diagnosis) rarely can be defended. Thus, not meeting the criteria may not be so special or important. Eliminating a problem behavior is quantitative in some sense (from some high rate to zero) but also is qualitative (from something to nothing) and that is a stark difference.

This index of clinically significant change is the easiest criterion to invoke and understand, but it is the least frequently used.

The reason is that changes in most problems of therapy (bipolar disorder, depression, anxiety, social withdrawal) are a matter of degree and hence decisions need to be made about whether the degree change is one that really makes a difference in the lives of the clients rather than completely eliminates a problem. Yet elimination of a problem is useful to mention because one can see better that a practical benefit can come from a very large and qualitative change. Yet, smaller changes that are not all or none can be important too.

Recovery. There has been interest in defining recovery from physical and mental health dysfunctions. One would think that this is easy, in part because we are used to things like the common cold in which we have it and then days later do not, i.e., are recovered. However, recovery is not so clear for many conditions (e.g., chronic medical disorders), including psychotherapy (e.g., for depression, anxiety).

A working definition of recovery has been sought that would be of use to researchers, mental health professionals, insurance and third-party payers, and policy makers. Our interest is in relation to evaluating outcome.

Recovery has been defined in the context of mental disorders and addictions as, "A process of change through which individuals improve their health and wellness, live a self-directed life, and strive to reach their full potential."

(Substance Abuse and Mental Health Services Administration [SAMHSA], 2011, quote from Web page, see full reference.) The definition includes functioning in different spheres, including health (living in a healthy way or managing one's condition), home (having a stable and safe place to live), purpose (e.g., having a purpose and being involved in meaningful activities), and community (e.g., having relationships with others and social networks with support, friendships, love, and hope). There are multiple facets of recovery as one can see and with that recovery is clearly dimensional rather than all or none and can vary by individual dimension (e.g., purpose, home) (SAMHSA, 2011; Whitley & Drake, 2010). There are measures available, but they are not frequently used in the context of psychotherapy (e.g., Andresen, Caputi, & Oades, 2006). The concept of recovery and an evaluation would be a useful addition for evaluating clinical significance of treatment. Among the reason is that it turns away from a sole symptom focus and draws attention to how one is doing in life.

Quality of Life. For a period spanning decades, there has been a call to use quality of life measures to evaluate clinical significance of the change and client functioning (Awad & Voruganti, 2012; Frisch, Cornell, Villanueva, & Retzlaff, 1992; Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). And contemporary research both in psychology and medicine uses quality of life as an index to complement direct assessment of change in symptoms and disease (Lerman, BGS, Gellish, & Vicini, 2012; Lindner, Andersson, Öst, & Carlbring, 2013). *Quality of life* refers to the clients' evaluation of how they are doing in several spheres of life.

For example, the Quality of Life Inventory, one of several available measures, includes multiple domains (e.g., health, philosophy of life, standard of living, work, recreation, love relationship, relationships with one's children, friendships, and others) (Frisch et al., 1992). In this measure and others, clients rate their satisfaction with their lives in that domain. As might be expected, ratings of quality of life are negatively correlated with psychological symptoms but are not redundant with that.

The construct overlaps with other indices of clinical significance we have already discussed, especially subjective evaluation and recovery, but is worth distinguishing. Among the reasons, quality of life is intuitively attractive and translates to what many people can readily understand.

Also, quality of life has been investigated much more broadly and beyond the context of dysfunction. The scope of research includes extensive use in medicine to evaluate the impact of treatment (e.g., surgery, medication) for a vast range of disorders. In addition, a measure of quality of life is easily added to an assessment battery and many measures are available.

12.6.5: Further Considerations Regarding Other Criteria

Qualitative Assessment. Qualitative measures focus on in-depth evaluation of individuals and look for themes that might capture that individual's experience. The advantage of qualitative assessment is that the focus is on the individual.

The definition of the benefits of treatment and whether they are important may require the individual's perspective (Hill et al., 2011). Also qualitative assessment focuses on dimensions that are not defined by the investigators or the assessment device. For example, in evaluating treatment outcome, standardized (quantitative and objective) measures are used. These do not allow entry into other domains than those included in the scale. In contrast, qualitative assessment usually is open ended with questions that have no fixed answers. Clients are allowed to tell their story and how they have changed or not. In the very few studies that have used qualitative assessment, it is clear that dimensions that clients report as important at the end of treatment (e.g., gaining more control in their lives, being able to care for others more, being more open emotionally) depart from the usual outcomes (e.g., symptom change) (e.g., Morris, 2005; Nilsson, Svensson, Sandell, & Clinton, 2007).

Qualitative assessment is a viable option for evaluating the impact of treatment (see McLeod, 2011).

Strength of the assessment is evaluating the clients' views of how therapy did and did not help and whether any changes make a difference and along what dimensions.

Yet, qualitative assessment is labor-intensive. In-depth interviews are required, and scoring of the interview data is not familiar to most researchers trained in the quantitative tradition. I mention the method here because of its clear relevance and utility in defining clinical significance. Of all measures, qualitative assessment gives greatest weight to the views of the clients and how they see the benefits (or limits) of the changes they have made.

Social Impact Measures. Another type of measure that helps to evaluate the clinical or applied importance of treatment outcomes is to see if measures of social impact are altered. Social impact measures refer to outcomes assessed in everyday life that are important to society at large.

For example, rates of arrest, truancy, driving while intoxicated, illness, hospitalization, and suicide are prime examples of social impact measures. Another measure is the use of health care services after treatment. Presumably a treatment with social impact would be one in which individuals needed to rely less often on other and possibly more expensive health care services. A reduction of visits to an emergency room or to medical or psychological services would be such measures. Such measures often are regarded by consumers of treatment (i.e., clients who seek treatment, insurance companies or employers who pay for treatment) as the bottom line. To the public at large and those who influence policy, social impact measures are often more meaningful and interpretable than the usual psychological measures. At the end of treatment, psychologists may become excited to show that changes were reflected in highly significant effects on the usual psychological measures (e.g., Hopkins Symptom Checklist, Beck Depression Inventory). However, what does this "really" mean? To the public, the effects are clearer if we can say that the effects of treatment are reflected in reduce absenteeism from work, fewer visits to the doctor for health problems, or fewer suicides.

Social impact measures have often been used in clinical and applied studies that track individuals longitudinally. For example, early school intervention (first and second grade) with disadvantaged elementary school students focused on a special classroom program (Good Behavior Game). Several studies have attested to the near-term effects in reducing disruptive behaviors. Later social impact measures were evaluated when the youth were aged 19–21 (Kellam, Reid, & Balster, 2008; Poduska & Bowes, 2010). The impact of the program was remarkable. Youth who were aggressive and disruptive in elementary school and who received the program years later showed reduced:

- Drug and alcohol abuse and dependence
- Cigarette smoking
- Use of special school services (e.g., for problems with behaviors, drugs, alcohol)
- Violent and criminal behavior
- Antisocial personality disorder
- Suicide ideation and attempt

Some of these measures (e.g., smoking, alcohol use) might be considered to be measures of other symptom domains, but they were related to the use of services, i.e., social impact measures. In passing, the Good Behavior Game intervention has been used in elementary through high school classrooms and in regular and special education classes and has strong evidence in its behalf (see Embry, 2002; Tingstrom, Tingstrom, Turner, & Wilczynski, 2006).

Social impact measures are clearly relevant for evaluating treatment. In many ways, the measures might be conceived as having social significance rather than clinical significance.

By social significance I mean measures that reflect indices especially important to society and that focus on more of a group-level outcome. We want to know if a treatment reduces arrest rates and homelessness of course. The social impact measures focus on rates of these for a group.

Clinical significance, in contrast, emphasizes the impact of the intervention on the individual client. The measures are not incompatible, and indeed one can look at individuals and measures as an outcome of the number of occasions in which they have had contact with law enforcement agencies or drug rehabilitation centers. In this case, the issue is how the social impact measure is used, to evaluate groups and/or individuals. Yet, the measures are valuable in part because they readily reflect domains of direct interest to many constituencies (e.g., policy makers).

12.6.6: Other Terms and Criteria Worth Knowing

I have covered the main measures used to evaluate clinical significance and lesser used indices that are just as defensible for use. There are closely related concepts important to know in clinical psychology that connect with clinical significance or importance of the change.

> First is the notion of impairment, sometimes also referred to as disability. *Impairment refers to the extent to which the individual's functioning in everyday life is impeded*.

Meeting role demands at home, work, or school; interacting prosocially and adaptively with others; and being restricted in the settings, situations, and experiences in which one can function can vary considerably among clients with a given disorder or problem. Impairment usually is a criterion that is central to defining a psychiatric disorder, i.e., do one's symptoms interfere with everyday functioning and also contribute to the likelihood of being referred for mental health services?

What do you think?

A measure of clinical significance might well be defined as meeting some cutoff or amount of improvement that is reflected in reduced impairment. Impairment as a concept is important to know because it is central to psychiatric diagnosis. I did not list this as a measure of clinical significance because of the overlap with recovery and quality of life that reflect the positive side of the concept, i.e., doing well in domains of everyday life.

> Second is the concept of *disability-adjusted life year* (DALY), which is a measure often used in epidemiological studies to evaluate the burden of disease as applied to dysfunctions of mental and physical health.

DALY refers to 1 year of a "healthy" life that is lost due to some impairment (e.g., depression, cancer) (see World Health Organization, 2013, for further information regarding how this is calculated). DALY provides a single estimate of the burden by combining both morbidity (disease or condition) and mortality (and lost years in the case of death or life expectancy) (see World Bank, 1993, for details).

For example, we know that mental disorders are more impairing than common chronic medical disorders, with particularly greater impairment in the domains of home, social, and close-relationship functioning (Druss et al., 2009). As a dramatic illustration, in 2004, the burden of depressive disorders (e.g., years of good health lost because of disability) was ranked third among the list of mental and physical diseases (World Federation for Mental Health, 2011). By 2030, depression is projected to be the number one cause of disability in the world, ahead of cardiovascular disease, traffic accidents, chronic pulmonary disease, and HIV/AIDS (WHO, 2008). Estimates like these are based on DALYs. Among the advantage is a common metric that is used to evaluate many different types of disease and social burdens.

DALY is not a measure commonly used in clinical psychology or to evaluate the therapy when clinical significance is an issue. Yet, it is important to be aware of the measure and to ponder its use and integration to connect with a broader literature (e.g., public health, psychiatry, epidemiology). Also, the measure might be considered to overlap with or be a type of social impact measure as highlighted previously. The emphasis is on group data rather than change of the individual.

> Both impairment and DALYs convey there is a broad concern in evaluating interventions and their impact beyond symptom change. There are now many ways of looking at impact both on the individual and society at large.

This broader concern underscores the importance of including measures that get closer to genuine impact on client functioning than the usual changes in objective measures most commonly used to evaluate treatment. Objective measures and change on standardized scales are wonderful, and some of my relatives even use them. The concern is not about those measures and their use but rather the limitations of what they can address, namely, genuine impact of treatment on the lives of individuals.

12.6.7: General Comments

I have mentioned the main measures of clinical significance used in clinical psychology as well as another set of measures that are options. I have not exhausted all of the possibilities (e.g., Patterson & Mausbach, 2010). Also, it is clear that among those I have presented there is overlap.

From the discussions, there are a few critical points to underscore:

1. Most treatment outcome studies that focus on psychiatric disorders or emotional, cognitive, and behavioral problems do not evaluate the clinical significance of change. Consequently, it is worthwhile to integrate at least one of the indices I have highlighted into one's research.

2. There is no standard measure of clinical significance routinely used in treatment studies or that is strongly preferred. Each of the measures has its own challenges. For example, using the normative range as a basis for clinical significance raises questions about who the proper comparison group is, whether that would vary by culture or other subject demographic variables. In the reliability of change measure, a dramatic change (about two standard deviations) could be partially due to statistical regression because clients who receive treatment usually are selected because of their extreme scores to qualify for the trial. And so on—each measure could be evaluated critically.
3. It is useful to use more than one index of clinically significant change. The reason is that the different indices yield different results and conclusions (Ronk, Hooke, & Page, 2012). Two or more indices provide a fuller picture of the scope and magnitude of the changes. This is a feasible suggestion because some of the indices (e.g., reliability change index) do not require new measures but rather how those measures are evaluated.
4. The most significant, a haunting concern remains with the most commonly used measures of clinical significance. We do not really know whether clients have been helped in a practical way (Blanton & Jaccard, 2006; Kazdin, 2006; Rutledge & Loh, 2004). For example, how a return to a normative range or showing a large change (reliable change index) on a psychological measure translates to functioning in everyday life is not well known or studied.

With limitations and nonstandardization of many measures, there is a larger point. When evaluating interventions that are designed to help people, whenever possible include some measure that goes beyond the usual questionnaires and other such measures.

> The question to address in the study, "What evidence can be provided that the intervention had personally significant impact on the lives of individuals who received the intervention?"

Personal significance, so to speak, has no $p < .05$ level the way statistical significance does. Also, the measure has no necessary relation to effect size. Rather personal significance is concerned with—have we helped you (client) with your definition of what that means rather than with my (researcher) definition of what that means? Once cast in this light, it is surprising perhaps how little empirical attention has been given to the topic and how few studies actually assess clinical significance.

One need not slavishly adhere to currently used measures of clinical significance, highlighted here. Indeed, some words of caution are in order for the commonly used measures. Measures of clinical significance are defined largely by researchers, with the exception of social impact measures. That is, the measures are operational definitions of what we consider reasonable bases for saying that there was clinically important impact. As paradoxical as it may sound, there is little or no evidence (depending on the measure) that what we as researchers call clinically significant makes any real difference in the lives of the persons who receive treatment.

> When we talk about clinically significant change, the assumption is that we want large change and one that really makes a difference. Yet, small effects of therapy—or making just a little difference—might still be very important and clinically important because of the difference that make in a person's life.

If treatment makes people a little better (e.g., a little less depressed or anxious, a little more confident, they drink or smoke less), this may be enough to have them enter parts of life (e.g., social events, relationships, work, new hobbies) that they would not otherwise do. Similarly, making a dysfunctional marriage a little better may have important impact on the couple and the individual spouses (e.g., deciding to remain married during a period of turbulence), even though on a measure of marital bliss, the couple still falls outside the normative range. Notwithstanding these considerations, the researcher is encouraged to include one or more measures of clinical significance in any intervention study. The purpose of the addition is to move beyond mere statistical significance to something more akin to personal significance.

12.7: Assessment during the Course of Treatment

12.7 Express how regular and periodic evaluation of client progress is carried out during the course of treatment

Controlled trials of treatment usually consist of evaluation of participants before treatment begins and then after treatment is completed (pre- and posttreatment assessment). Not many investigators assess clients during or over the course of treatment. Ongoing assessment over the course of treatment consists of regular or periodic (e.g., every session, every other session) evaluation of client progress. Consider research and clinical advantages of ongoing assessments in the context of intervention research.

12.7.1: Evaluating Mediators of Change

An important research focus in intervention research is to understand how and why changes occur over the course of treatment. We want to know why treatment led to change

so that we can identify the critical ingredients and maximize change. Virtually all therapies have a conceptual view about why people get better in treatment, although rarely is there any supportive evidence for that interpretation. Tests of mediation are studied in an effort to identify the processes involved that underlie, account for, mediate, and cause therapeutic change.

A study might propose that changes in cognitions are the reason patients improve, i.e., is the mediator.[2] At the end of treatment, the investigator may show that symptoms changed and, as hypothesized, cognitions changed too. If both symptoms and cognitions have changed at the end of treatment, it is not possible to state that change in one preceded the other. Perhaps they both changed simultaneously, perhaps symptom change caused cognitive change or vice versa, and changes in symptoms were caused by some other influence (e.g., expectations for improvement). As noted previously, among the many requirements of a causal role and mediation is establishing the time line.

> The study of mediators of change in therapy requires assessment during the course of treatment. Mediators (e.g., changes in cognitive processes) usually are measured at a fixed and predetermined point in time or let us say even two points in time.

It might well be that for all persons in the study the proposed mediator in fact accounted for therapeutic change. Even if all patients change on the basis of the identical mediator, the timing and patterns of change may vary (e.g., Stulz & Lutz, 2007). It is now well documented that some patients make rapid changes quite early in treatment (referred to as sudden therapeutic gains), as has been shown, for example, in clients treated for depression or anxiety (e.g., Aderka, Nickerson, Bøe, & Hofmann, 2012; Tang, DeRubeis, Beberman, & Pham, 2005). That finding alone, now well replicated, suggests that mechanisms operating to produce change vary in when they operate. In addition, brain activation (fMRI) and symptom change vary in magnitude and relation at different points over the course of treatment and are not the same for all individuals (e.g., Schiepek et al., 2009). In short, the course of change and hence the course of the processes underlying change vary among individuals. Assessment of the mechanism at any one or two points in a study may not capture when change in the mechanism has occurred for each individual.

> A so-called negative finding ("no relation" between mediator and outcome) may result because the mediator was not assessed at the optimum point for each participant in the study.

As an illustration, a great deal of research has focused on the relationship between the therapist and client forms during the course of treatment. Literally thousands of studies have evaluated the contribution of the therapeutic relationship to therapy process and outcome (e.g., Horvath & Bedi, 2002; Norcross, 2011). The extensive research is understandable, given the proposed role of the relationship in many different approaches to therapy (Prochaska & Norcross, 2010) and evidence that the relationship influences treatment across a broad range of clinical dysfunctions including severe mental illness (McCabe & Priebe, 2004). Among the many facets of the relationship, the therapeutic alliance is most frequently studied (Castonguay, Constantino, Boswell, & Kraus, 2011). Alliance refers to the quality and nature of the client–therapist relationship, the collaborative interaction, and the bond or personal attachment that emerges during treatment. A high-quality therapeutic alliance is associated with higher rates of treatment completion, greater compliance with treatment demands, and greater therapeutic change (e.g., Norcross, 2011; Orlinsky, Rønnestad, & Willutzki, 2004).

There are many measures and views of alliance, and their differences need not detain us. One view has been that alliance leads to (mediates, causes) therapeutic change. That is, during the course of treatment an alliance develops, and if this is a good alliance, this predicts improvement in symptoms. To study this requires that one looks at alliance during treatment.

12.7.2: More Information on Evaluating Mediators of Change

> *Assessment usually is assessed by administering a standardized measure of quality of the alliance (self- and therapist-report measures) one or two occasions during treatment and relating these (e.g., correlation) to improvements in therapy.*

A well-established finding is that the stronger the therapeutic alliance during treatment, the greater the therapeutic change by the end of treatment (Crits-Christoph, Gibbons, & Mukherjee, 2013; Norcross, 2011), although the magnitude of the relation is rather small (r = .27) (see Horvath, Del Re, Flückiger, & Symonds, 2011, for a meta-analytic review).

Alliance usually is assessed at one or two fixed points during treatment. An implicit assumption is that one or two particular points adequately sample the change in the mediator for all participants. This assumption is almost certainly false in light of individual variation in patterns of change, as noted previously. A challenge for research is ensuring that one can evaluate mechanisms and change that may vary in course among individuals. To do that, we need ongoing assessment of each participant, i.e., assessment on multiple occasions (e.g., on a session-by-session basis in group studies) (e.g., Lutz, Stulz, & Kock, 2009). That assessment allows the ability to examine the mediator–outcome relation for each participant and at different points in time.

Single-case designs that rely on continuous (ongoing assessment) or group designs with multiple assessment occasions can reveal the individual patterns. Ongoing assessment can establish the time line of proposed mediator and change but also elaborate the nature of the relationship. For example, assessing alliance and symptom changes at each session brings to bear the strength of identifying the time line of change and the relations between putative mediators and outcomes. Recent studies in which session-by-session assessment was completed revealed that alliance predicts symptom improvement but symptom improvement also predicts alliance (Lutz et al., 2013; Marker, Comer, Abramova, & Kendall, 2013). This is a reciprocal determinism that would not be detected or detectable in single occasion assessment of a mediator at a fixed period during the course of a study.

More generally, the studies of mediators would profit enormously from ongoing assessment over the course of the intervention. This is a stark departure from the vast majority of research on mediation that uses only one occasion to assess the mediator over the course of treatment. If all subjects changed due to the mediator, this would not be picked up in a study that only assessed the mediator on one or two occasions just because individuals vary in rate and course of change. I noted previously that the magnitude of the relationship between alliance and therapeutic change is small. This might be accurate, but we cannot know because the vast majority of studies looked at this relation with fixed assessments and only picked up the relation of alliance and change restricted to these assessment occasions. The data were averaged across many participants at these one or two assessment points. The few assessment occasions and combining data from multiple patients would obscure and likely hide or underestimate mediator–outcome relations. The relation may be stronger or different with ongoing assessment that permitted evaluation of the pattern of change for each participant.

12.7.3: Improving Patient Care in Research and Clinical Practice

Apart from understanding mechanisms, ongoing assessment can improve research on treatments that will have impact on clinical care and in addition improve the quality of clinical practice. Let us begin with the research point. The vast majority of therapy studies provide a fixed number of treatment sessions with minor opportunities for flexibility (e.g., adding one or a few sessions).

What can you say from pre- to post-assessment?

At the end of the post-assessment, we can make such statements as this client improved, did not change much, or became worse. Often we would like to know what is happening during treatment, i.e., how the patient is functioning while treatment is in process and to use that information to make decisions about treatment and how it ought to be optimally applied. Ongoing assessment can inform how to carry out treatment in clinical practice.

Presumably some clients require more and others less than the fixed regimen the investigator selects somewhat arbitrarily. And to understand treatment better, it would be useful to know the likely amount of treatment or scope of treatment experiences will achieve the likely level of gains and whether that amount varies systematically as a function of some other variable(s). As I mentioned, some patients make rapid changes quite early in treatment. Other clients may not make expected changes and are unresponsive even to extended treatment. And, of course, there are the gradations in between and the cases in which change occurs in some areas of functioning but not in others or at different rates among the various areas.

> It is important to monitor treatment effects in an ongoing way to make decisions about continuing, altering, or terminating treatment on the basis of how well the client is doing.

Ongoing assessment would provide a better idea about treatment progress, number of optimal sessions usually required for change, and how different points in treatment might predict poor or good therapeutic outcomes. From research we want more than knowing if treatment is effective. We also want to know how treatment is optimally applied (e.g., amount of treatment) and whether there are decision points during treatment that make it likely that further treatment will not be effective and likely that some other intervention should be applied. Pre- and post-assessment in randomized controlled trials of therapy are not clinical-application friendly because we need information along the way, client progress, and interim outcomes, to guide decision making. The discussion of continuous assessment in the chapter on single-case design conveyed the advantages of ongoing assessment. Group designs too could integrate ongoing, regular assessments over the course of treatment to better understand process and change.

12.7.4: More Information on Improving Patient Care in Research

Turning to clinical practice further elaborates the benefits of ongoing assessment. Most patient care (psychotherapy, mental health services for individuals or groups) in clinical practice is not evaluated systematically. Clinical judgment is relied on without the benefit of complementary systematic measures that could promote more informed decision making to the benefit of the client. Ongoing clinical assessment arguably is as if not more important in clinical work

than it is in research. Real people are involved in clinical work, they are seeking help or untoward and often debilitating conditions, and we are trying to make them better. This is slightly different from a controlled trial where we would very much like people to get better but are testing hypotheses under controlled conditions and have noted in consent forms this is a research project and participants may not improve. In clinical work, even the best evidence-based treatment may not work or occasionally make a patient worse. Systematic assessment at the very least would be an excellent supplement to clinical judgment.

> Ongoing assessment in research and practice is quite feasible. For example, at the beginning of each treatment session, a few minutes (e.g., in the waiting room or treatment session) might be used to have the client complete a brief measure to evaluate if there are any changes in key domains of interest.

Data obtained every other session or on some days during the week between sessions are other options to achieve the same goal, namely, to monitor progress in therapy. User-friendly (brief, straightforward) measures have been developed that are readily available for clinical use. One example is the Outcome Questionnaire, with multiple forms and with versions for therapy with adults as well as children. One variation is the Outcome Questionnaire 45 (OQ-45), which is a self-report measure designed to evaluate client progress (e.g., weekly) over the course of treatment and at termination (see Lambert, Hansen, & Finch, 2001; Lambert et al., 2003). The measure requires approximately 5 minutes to complete and provides information on four domains of functioning:

- Symptoms of psychological disturbance (primarily depression and anxiety)
- Interpersonal problems
- Social role functioning (e.g., problems at work)
- Quality of life (e.g., facets of life satisfaction)

The measure has been thoroughly evaluated with thousands of patients and with individuals of different cultures and nationalities and shown to predict response to treatment among individual cases. Moreover when the measure is used to provide feedback to therapists, the patient outcomes are improved by decreasing the number of patients who deteriorate with therapy (e.g., Lambert & Shimokawa, 2011; Shimokawa, Lambert, & Smart, 2010). In treatment as usual, ongoing assessment and feedback lead to better outcomes than treatment as usual without the feedback (Simon, Lambert, Harris, Busath, & Vazquez, 2012). The absence of systematic evaluation routinely in clinical practice is a critical issue independent of the issue of whether an EBT is the core treatment. It is clear now that both the interventions that are used and how client progress is monitored can influence therapeutic change.

The OQ represents a family of measures with various versions and short forms. Yet, there are many other opportunities for ongoing assessment. Assessment of psychological symptoms and many other facets of functioning (e.g., mood, arousal, stress, cognitive processes, use of self-regulation strategies) can be done in real time via smartphones, tablets, and smart watches with "apps." These measures allow assessment in everyday life as the clients negotiate their regular routines. Such assessments can be completed in an ongoing way and even be sent and stored (encrypted for privacy) to the clinician's computer and database. There is nothing here that is new technology. Alternatively, the paper-and-pencil measure in the waiting room (e.g., OQ 45) will serve just as well so that technological impediments are not a reason to omit ongoing assessment.

The hiatus between research and clinical practice continues to be an issue in clinical psychology (Kazdin, 2008b). Much of the discussion is about evidence-based treatments and why they are not routinely used in clinical practice. It is true that most treatments used in clinical practice are not evidence based. When they are evidence based, their effectiveness is not up to the level attained in research (Weisz, Ng, & Bearman, 2014). Yet, perhaps even more significant is the absence of systematic evaluation. This is a case where systematic assessment appears to be in the best interests of the clients.

12.7.5: General Comments

Ongoing assessment is not better or worse than assessment on one or two occasions (e.g., pre- and post-intervention). As any methodological practice, the issue is what one wants to conclude from a research project. *Ongoing assessment provides information about the course of change*; and when this is evaluated on an individual by an individual basis, the picture of how change occurs is likely to be very different from the data obtained on one or two assessment occasions during treatment and collapsing (combining) the data across subjects. Also, investigation of mediators or mechanisms of change are likely to require ongoing assessment because there is no reason to believe that a mediator will or will not operate by some arbitrary session or two when a fixed assessment was conducted.

Ongoing assessment and its advantages were discussed briefly in the context of clinical care. Clinical practice is beyond the scope of this text. Yet it was mentioned because research methods, in this case systematic assessment in an ongoing way, could be used and have been in studies cited to improve patient care and outcomes of treatment. Ongoing assessments may be a way to not only help patients but also use data from clinical settings to enhance our understanding of treatment.

Summary and Conclusions: Special Topics of Assessment

Three assessment topics served as the basis of the present chapter:

- Assessing the impact of experimental manipulations
- Assessing the clinical significance of change
- Use of ongoing assessment

Each has direct implications for what one can conclude from a study. Assessing the impact of the experimental manipulation is a check on the independent variable and how it was manipulated or received on the part of the participant. If the independent variable has its intended or predicted effects on the dependent measure, this is usually sufficient evidence that the intervention was adequately manipulated. It is desirable to have additional information to assess whether the independent variable was adequately manipulated. A check on the manipulation can be obtained by determining whether the subjects were affected by the particular changes in conditions or whether the procedures to which they were exposed were executed properly. This check, distinct from performance on the dependent variables, provides some assurance about the adequacy of the experimental test.

Occasionally, interpretive problems may arise if there are discrepancies between the information provided by check on the manipulation and the dependent variables. However, as a general rule, assessing the adequacy with which the independent variable is manipulated can be extremely useful both for interpreting the results of a particular experiment and for proceeding to subsequent experiments. Interpretive problems that can arise in experiments can be attenuated in advance by conducting pilot work to explore different ways to manipulate the conditions of interest. Assessment of the experimental manipulation can ensure that the independent variable receives the most potent empirical test.

Assessment of clinical significance of change was the second topic of the chapter. Clinical significance reflects the concern about the importance of therapeutic changes that were achieved. Statistical significance on the usual measures does not necessarily mean that the treatment has had any impact in ways that are of practical value in the lives of individuals. Several indices of clinical significance have been used. The most common are showing that after treatment the symptoms have returned within a normative range, that very marked changes have been made (e.g., approximate two standard deviations), that diagnostic criteria used to select participants for treatment are no longer met, and that subjective evaluations show that clients or those with whom they interact see the changes as important. Other means were mentioned as well (e.g., recovery, quality of life). No single method is used extensively, and whether the most commonly used measures reflect client experience in ways that have made a difference in their life is not well established at all.

The third assessment topic was the value of ongoing assessment of treatment in intervention studies. The usual assessment consists of pre- and posttreatment assessment to evaluate improvement. If the goal is to understand the course or mediators of treatment, ongoing assessment (on multiple occasions over the course of treatment) can be very valuable. The course of change both of mediators and symptoms are likely to vary among individuals. Currently when studies evaluate mediators, one or two assessment occasions are used during treatment to assess the mediator (e.g., cognitive processes, alliance). Fixed and few assessment occasions during treatment are likely to miss the patterns of changes and to misrepresent mediator and outcome relations. Ongoing assessment was also discussed briefly in the context of clinical care. There are user-friendly, brief, and well-validated measures that could be used in clinical practice. When used they have been shown to improve patient outcomes. A major contribution of research and research methodology to patient care is the development and validation of measures. Perhaps someday this will actually influence how clinical care is conducted.

Critical Thinking Questions

1. Identify some strengths and weaknesses of checking on the experimental manipulation.
2. What are your views of the best or at least highly desirable ways of measuring clinical significance? Identify two. Why do you think they are important?
3. What are some advantages of using ongoing measures over the course of treatment for research (understanding how therapy works)? And for clinical practice (helping individual clients)?

Chapter 12 Quiz: Special Topics of Assessment

Chapter 13
Null Hypothesis Significance Testing

Learning Objectives

13.1 Report the historical development of statistical tools

13.2 Review the utility of the alpha in statistics

13.3 Analyze the significance of a strong power in statistical tests

13.4 Determine the different ways of increasing power in statistical tests

13.5 Express why the design stage of any robust statistical test needs to account for data analysis

13.6 Report the arguments against the use of the null hypothesis significance testing in statistical analysis

13.7 Examine the Bayesian data analyses as an alternative to the null hypothesis significance testing in statistical analysis

Empirical research in the natural, biological, and social sciences is based primarily on the model referred to as *null hypothesis significance testing*, frequently abbreviated as NHST. This is the model of research within the quantitative tradition that begins by posing that the experimental manipulation will have no effect, i.e., the null hypothesis. That is, we are comparing various groups or experimental conditions, and the analyses we conduct begin with the view that there is no difference among the various conditions. We need a strong reason or consistent way of deciding when to reject that hypothesis.

> The null hypothesis is rejected or accepted based on whether the differences between groups are statistically significant by a predetermined criterion (typically $p \leq .05$).

Psychological research is very much dominated by NHST—in many ways, it is like the water in which fish swim—so pervasive we do not think about it very much. Consequently mastering research methodology, designing one's own individual studies, and consuming (reading) the research of others require understanding central features of NHST. The overall approach and concerns about its limitations are important to know generally in part because they can greatly influence how a study is planned long before any data are collected as well as how the data are presented and analyzed once the data finally are collected. Understanding the approach in more detail, key concerns, and how they can have impact on one's study can greatly improve the quality and options in one's own research. As important, in reading the research of others it is critical to understand what can and cannot be said given the key features of NHST.

This chapter focuses on statistical evaluation of the data. In this chapter, I discuss the null hypothesis testing approach and the concerns about the approach, key concepts of the approach that are pertinent to consider at the planning stage of the study, and additional options about hypothesis testing in addition to statistical significance testing.

13.1: Significance Tests and the Null Hypothesis

13.1 Report the historical development of statistical tools

Once upon a time there was no statistical testing and no statistical evaluation. Actually, this statement probably reflects my poor scholarship. No doubt one of the first uses of statistical tests, like so many other firsts, can be traced to the ancient Greeks and most likely to Aristotle. The first statistical evaluation emerged when Aristotle's mother played the money game with him when he was 4 years old.

His mother held out two closed hands (fists) and said in a playful way, "Ari, which hand holds *more* drachma (Greek currency before the Euro); if you guess correctly, you can keep the money in that hand?" Ari replied, "Trick question mom, although one hand has three coins and the other has one coin, the two numbers are not really different statistically speaking, that is. I'm guessing that comparing the number of coins in each hand would not be statistically significant at the $p < .05$ level. Mom, we could not reject the null hypothesis (number of coins in your hands are no different from each other). Nice try mom—you almost had me." Aristotle's mom, no slouch herself (e.g., after all, she spoke ancient Greek fluently), quickly replied, "If they are not different, then let me give you what is in this hand!" at which point she handed him the one coin. Aristotle expressed grave dissatisfaction as he reached out with his open hand to receive the one coin. Yet, from this moment on he grasped his mother's lesson, namely, that one can accept the null hypothesis—no difference—when it is really not wise to do so—that is the first Type II error ever recorded (accepting "no difference" when in the world there really is a difference).

In any case, invoking statistical significance as a criterion for decision making was a major contribution to science, for which we thank Aristotle. Showing that even when there is no statistically significant difference, there may be a real and important difference and this also is a major contribution, for which we thank his mom. Apparently, Ari's dad walked in and overheard but did not understand all of this and just complained that the whole conversation was Greek to him.

Moving forward a bit in time and to nonfictional history, it is useful to stop in the 1920s and 1930s. During this period, statisticians devised practices that dominate current statistical methods of evaluation in psychology, and indeed in the sciences in general (Fisher, 1925; Neyman & Pearson, 1928). The practices include posing a null hypothesis (an assumption that there are no differences between groups in our study) and using tests of significance to determine whether the difference obtained in the sample is of a sufficient magnitude to reject this hypothesis.

> That is, what might be a reasonable or reasonably stringent criterion to reject that no-difference hypothesis?

Of course, we "really" believe and want group differences, but we begin with the assumption that unless there is compelling evidence, we shall take the stand that there are no differences.

> A goal of statistical evaluation is to provide an objective or at least agreed-upon criterion (e.g., significance levels) to decide whether the results we obtained are sufficiently compelling to reject this no-difference hypothesis.

After all there are likely to be some differences between groups (e.g., an experimental and control group) on whatever measures we used to evaluate the results. Means would rarely be identical for any two (or more) groups on any measure. Indeed, statisticians are fond of noting that the means between (or among) groups are always different. Strong, cryptic, and clear quotes convey the sentiment: "It is foolish to ask 'Are the effects of A and B different?' They are" (Tukey, 1991, p. 100) and "So if the null hypothesis is always false, what's the big deal about rejecting it?" (Cohen, 1990, p. 1308). This point is important to keep in mind as we move forward on the interpretation and limitations of searching for statistical significance.

In fact, in contemporary research statistical tests are used to decide whether the differences in a particular study are likely to reflect a real or reliable difference in the world and are not likely to be due to a fluke (coincidence, chance) of whom we sampled as subjects. Statistical evaluation examines whether groups differing on a particular independent variable (e.g., different conditions) can be distinguished statistically on the dependent measure(s). Statistical evaluation consists of applying a test to assess whether the difference obtained on the dependent measure is likely to have occurred by "chance." Typically, a level of confidence (e.g., .05 or .01) is selected as the criterion for determining whether the results are *statistically significant*.

> A statistically significant effect indicates that the probability level is equal to or below the level of confidence selected, for example, $p \leq .05$; if the experiment were completed 100 (or better an infinite number of) times, a difference of that magnitude found on the dependent variable would be likely to occur only 5% of times on a purely chance basis.

So in our study, we plucked a set of subjects (a sample) at a particular point in time and did our study. It is possible that the sample and our finding do not represent what would be evident if we drew another sample from that population.

Think of it this way, if we flipped a coin (heads, tails) an infinite number of times, with an unbiased coin, there would be 50% heads and 50% tails. That is the probability of each in the entire population of all possible coin flips. Now let us say we just do 10 coin flips. It is possible albeit rare that our 10 flips of that same coin revealed 9 heads and 1 tail. What we obtain in any sample from the population may not truly reflect the population values. This is pretty much what null hypothesis statistical testing is about. In our study, if we sampled the population over and over again, we might find a "chance" finding 5% of the time (if $p \leq .05$ is our criterion for statistical significance). If the probability obtained in the study is lower than .05, most researchers would reject the null hypothesis and concede that group differences reflect a genuine relation between the independent and dependent variables.

To state that a relation in an experiment is statistically significant does not mean that there is necessarily a

genuine effect, i.e., a relationship really exists between the variables studied. Even a statistically significant difference could be the result of a chance event because of sampling of subjects and other factors. In any particular study, chance can never be completely ruled out as a rival explanation of the results. Nevertheless, by tradition, researchers have agreed that when the probability yielded by a statistical test is as low as .05 or .01, it is reasonable to conclude that a relation between the independent and dependent variables exists. As you can see, the criterion is a decision-making rule rather than a statement about reality, i.e., what is and is not really different in the world.

13.1.1: More Information on Significance Tests

Essentially, statistical evaluation provides a criterion to separate *probably veridical* from *possibly chance* effects. Statistical evaluation provides consistent criteria for determining whether an effect is to be considered veridical. This advantage is critically important. We lose sight of this advantage as researchers because we are sequestered from nonresearch-based influences and advocacy where the cannons of research are largely neglected. Media advertising claims about "effective treatment," for example, for losing weight, reducing cigarette smoking, or exercising to develop that Greek-sculptured and statuesque body are rarely based upon experimental methods and statistical evaluation. Testimonials by proponents of the techniques or those who may have participated in the programs serve as the basis for evaluation. Also, a nonscientific term "clinical evidence" usually is added to opinion and testimonials to give the aura or imprimatur of science. Needless to say, clinical evidence is not a recognizable research term in statistics and methodology—not even in the most comprehensive glossaries of terms (e.g., Zedeck, 2014). "Clinical evidence" is a made for TV (radio, Web) type term that is intended to imply "empirical research," "controlled studies," or even "randomized controlled clinical trials." The term means something like, "I have seen a lot of cases" or even worse, "In the opinion of the company selling the product and they made me use the term and also wear this white lab coat for emphasis."

There is another side.

> *Statistical significance is required in part because it is not otherwise clear in many or indeed most situations whether effects are beyond the differences or variations that would be evident by chance.*

Yet, clearly there are some situations in which statistical evaluation is not needed. I mentioned marked changes in the chapter on single-case experimental designs where nonstatistical data evaluation criteria are invoked. Indeed, the visual inspection criteria often used with these designs are intended to serve as a gross filter that allows only strong effects to be considered as veridical. In any type of research, whether single-case or group, very dramatic changes might be so stark that there is no question that something important, reliable, and veridical took place, the type of changes referred to as "slam bang effects" (Mosteller, 2010, p. 227).

A classic example in the 1700s was evident in the treatment of scurvy, a disease that results from a deficiency in vitamin C. Sailors and others in past times would contract the disease, become weak, have many lesions, and eventually die, and this occurred in large numbers while sailors were at sea. Among the reasons was that long voyages could not keep (store) fresh fruits and vegetables. Before the etiology was understood, a doctor (James Lind, 1716–1794) in the British Navy explored the effects of various interventions (e.g., special diet, drinking sea water, and eating citrus fruit like lemons and oranges, which of course are rich in vitamin C) among sailors ill with scurvy (Lind, 1753). This work involved few subjects and no statistical evaluation and included all sorts of methodological "problems" (little control of threats to validity) but the effects were stark. Individuals who were required to eat citrus fruit became better immediately (first couple of days) and could return to duties on their ship while their peers with the disease remained on the brink of death—"slam bang effects" that eventually eradicated the disease among sailors.

Clarity of the finding may have to do both the extent of impact of the intervention, the confidence one can place in the outcome measure (illness and death), and the clear predictable course without an intervention. And with the right combination, we do not need statistical tests. For example, most of us would be persuaded if three individuals who were terminally ill continued to live after a special treatment, whereas three others who did not receive the treatment died. We are likely to be persuaded by this demonstration in part because of the reliability, validity, and importance of the dependent measure (death), the virtual certainty of the predicted outcome without treatment, and the vast differences in the outcomes between expected versus obtained effects on the rate of death, even though as always we want replication of the demonstration. Most situations from which we wish to draw inferences do not show such "slam bang effects" or stark qualitative differences (dying vs. not dying). Even when such qualitative differences are evident, they are not likely to be evident for everyone in some intervention group (100% of participants) and none (0%) in the other groups.

> More likely than not, differences will require use of some criterion to decide whether the results, differences, or changes within or between groups are likely to be due to chance or random fluctuations. Statistical significance is designed to serve this purpose.

Endorsement of statistical evaluation does not mean that statistics provide "the answer" or "real truth." Statistical evaluation is subject to all sorts of abuses, ambiguities, misinterpretation, and subjectivity. For example, with statistical evaluation, there are many decision points in terms of the tests that are to be used, the criteria for statistical significance (e.g., $p < .05$, .01), whether there is control for the number of statistical tests carried out, and tacit assumptions and default criteria in the statistical tests themselves (e.g., for including or excluding a variable, for identifying the factor structure of a measure). These decision points often have no formal or agreed-upon rules or sources of justification but allow variation in analyzing and presenting one's data.[1] Often these ambiguities and decisions can be made explicit, studied, and understood and importantly checked by others. The explicitness of statistical procedures helps us raise questions and understand the limits of the conclusions.

13.2: Critical Concepts and Strategies in Significance Testing

13.2 Review the utility of the alpha in statistics

Significance level or alpha is well known as a criterion for decision making in statistical data evaluation. Tradition has led us to use an alpha of $p < .05$ and .01 for decision making.[2] Will the results of my experiment be statistically significant, or more carefully stated, will the differences between groups be statistically significant, if there are differences in the world? Among the determinants of the answer is the number of subjects per group in the study. As I have noted already, we can assume that groups will never (well, hardly ever) have identical means on the dependent measures, due simply to normal fluctuations and sampling differences. If group means are always numerically different, we have a bit of a problem.

13.2.1: Significance Level (alpha)

I mentioned before that statisticians are fond of noting that "groups are always different." The second part of that follows, namely, that "statistical significance" is a function or measure of sample size. That is, the larger the sample size, the smaller the group differences needed for statistical significance (alpha) for a given level of confidence. Stated another way, a given difference between two groups will gradually approach and eventually attain statistical significance as the size of the sample increases. Indeed, statistical significance is virtually assured if a large number of subjects are used.

For example, correlation coefficient (Pearson product-moment correlation or r) represents the linear association (relation) between two variables. A correlation can range from –1.00 to +1.00. A correlation of 0 or anywhere thereabouts means the variables are not linearly related. Yet $r = .01$ (i.e., a correlation of essentially 0) would be statistically significant at the $p < .05$ level with a sample of 40,000. But whoever has a sample this large? More often that one would imagine. For example, when psychological studies are conducted in the military (e.g., study suicide or traumatic injury), large-scale testing encompasses thousands of subjects—sometimes hundreds of thousands. In such circumstances, statistical significance is virtually assured. Large sample sizes make small, trivial, and chance differences more likely to lead to the conclusion that the results are "statistically significantly." The situation of large sample sizes is changing in light of some broader movements in science to pool the raw data from many individual studies and also to develop extremely large databases (referred to as big data).

Most psychology studies still include relatively small sample sizes (e.g., 10–50 participants for each group). Because statistical significance depends rather heavily sample size, we are likely to get varied effects—significant sometimes and not significant another because of that size—when we study the same phenomena in the same way! We want to know what variables are genuinely related to each other, how, and why. In obtaining this knowledge, we do not want our findings to wander in and out of zone of statistical significance because some samples are larger than others. Clearly, more information is needed than statistical significance. Yet in planning a study, the initial point to note is not to skimp on sample size.

13.3: Power

13.3 Analyze the significance of a strong power in statistical tests

It is important to revisit and discuss the issue of power because weak power is the Achilles' heel of psychological research. That is, if we are going to use tests of statistical significance to evaluate our results, it is critical to ensure that there is a strong chance (adequate power) of finding a difference when one in fact exists.

13.3.1: The Power Problem

The level of power that is "strong" is not derived mathematically. As with the level of confidence (alpha), the decision is based on convention about the margin of protection one should have against accepting the null hypothesis when in fact it is false (beta). A convention suggested decades ago has continued to be generally accepted, namely, that minimum power in a study ought to be .80 (Cohen, 1965). Power of .80 can be explained in statistical terms

(e.g., beta = .2 and 1-beta or power = .8), but there is a better way to state all of this. If power is .80 in a study, this means that the investigator's chance of detecting a difference in his or her study is 4 out of 5 (80%), if there is a real difference in the population.

> Remember power is the likelihood of finding a difference, but only the likelihood of finding a difference if there really is a difference in the population.

Power ≥ .80 is a widely accepted criterion within psychological science (e.g., Murphy, Myron, & Wolach, 2009), but certainly higher levels are desirable. An investigator is not likely to repeat any given experiment in exactly the same way, so increasing the "odds" of identifying a true effect has no inherent limit such as .8. Having a high level of power in a study is like exercise, green vegetables, and dental floss—all of them are good for us and serve our own interests and probably more than our usual amount is better.

Reviews of research within many different areas of psychology and other fields as well have shown that most studies have insufficient power to detect differences. Early evaluations of the power of research over 60 years ago to current analyses and many analyses in between, studies as a rule do not have sufficient power to detect small and medium effects (Cohen, 1962; Maxwell, 2004; Schutz, Je, Richards, & Choueiri, 2012). In one recent analysis of several areas of psychology, mean power was .35, quite below the recommended value (Bakker, van Dijk, & Wicherts, 2012). Moreover, repeated exhortations about the problem and consistently clear recommendations to rectify the problem have had little impact on research. (The value of the future exhortations in relation to clinical psychology has been in showing that insight and awareness into a problem and nagging people [investigators] are not very potent interventions for changing what people do.)

I also mentioned previously that power is not an esoteric concept or methodological annoyance. It can be a life or death matter when diagnostic or treatment options for diseases could not produce statistically effects because of weak power. Were the treatments really no different or not helpful? In clinical psychology and psychiatry, evidence-based treatments tend to be more effective than the usual run of the mill treatments conducted in clinics but not by a lot and not all of the time (Weisz et al., 2013).

> Could low power explain many of these effects or maybe this is just how things really are?

In many treatment and prevention studies, the conclusion is reached that two or more interventions did not differ from each other. No differences or support of the null hypothesis often is interpreted to mean that many treatments are probably equally effective. This may or may not be true. We cannot really tell because weak power is a plausible rival interpretation of the absence of differences. The inability to detect differences when they are there could be catastrophic as, for example, when comparing side effects (e.g., heart attack) of a medication across treatment and placebo controls groups and the treatment groups shows more of the side effects but they do not rise to the level of statistical significance (see Zilliak & McCloskey, 2008).

Low power is a huge threat to data-evaluation validity well beyond the contexts of treatment, and one that is critical to address explicitly in designing an experiment.

> Students often begin by asking advisors, "So how many subjects will I need for this study?"

The numerical answer can be estimated, as I note below, but the non-numerical answer usually is, "many more than you might think." When designing research, ensuring sufficient power is a huge consideration; when reading the research of others, recall that most studies are underpowered so look at that feature in evaluating the study.

As an all too frequent anecdote (you may have the same story), I just read an interesting article of a study on anxiety disorders and special characteristics that individuals with social anxiety may have. (Permit me to be vague on this not to indict author here.) Many comparisons across many measures were not significant. That is not a problem by itself at all. I will elaborate later on how and when "no difference" or negative effects can be interesting and important. (Also, after my dissertation, I came to "love" no differences.) Yet in this study, N (total subjects) for two groups was about 38. The results are very difficult to interpret because it is likely that any real effects would not be detected because of weak power.

> Is there really "no difference" in the comparisons of interest, *or* is the study too underpowered?

Many readers of the article might conclude that there are no differences; many researchers may not pursue the area further because the study suggests there is nothing there (no differences). The point is that weak power is not only a problem for an individual study but can greatly mislead. We believe the topic was studied and we have a reasonable answer, when in fact we do not.

13.3.2: Relation to Alpha, Effect Size, and Sample Size

Four different concepts of statistical inference have been discussed at varying points, including:

- The criterion for statistical significance (alpha)
- Effect size (ES)
- Sample size Power[3]

These concepts are interrelated in the sense that when three of these are specified, the remaining one can be determined. Their interrelations are critical in that they permit one to consider all sorts of options in an experiment, such as the level of power (given a specific level of alpha, ES, and a fixed N), what ES is needed (if alpha, power, and sample size are predetermined), and so on. The most frequent use of this information is to decide how many subjects to include in a study at the planning stage. Thus, to identify our sample size, we need to make decisions to fix the other three parameters:

- Alpha
- Power
- ES

At this point, let us adopt alpha of .05 to adhere slavishly to tradition. As for level of power, we also might follow convention and design our study to have power of .80. Now we must estimate ES. How can we possibly do this because the ES formula requires us to know the difference between the groups on the dependent variables of interest and the standard deviation ($ES = [m_1 - m_2]/s$)?

Actually, in many areas of research, ES has been studied. The secondary analysis procedure, referred to as *meta-analysis*, has been used extensively for evaluating research. Meta-analyses provide estimates of ESs for research in a given area. The ES is used as a common metric to combine studies using different dependent variables. We can consult such analyses to identify likely ESs for the study we propose to undertake. For example, one can readily search database (e.g., on the Web) that covers scholarly research in an area and seek "effect size for" and then name the intervention or area. As an illustration, a quick search of "effect size for mindfulness for depression" identified a meta-analytic review noting ES of .95 and .97 for patients with diagnoses of anxiety and mood disorders, respectively (Hofmann, Sawyer, Witt, & Oh, 2010). So one can readily search to find approximations of ESs before embarking on the design of a study.

If meta-analyses are unavailable, there may be one or more studies that can be found in the journals that have compared the conditions (groups) of interest or that used the measures (dependent variables) in a related way. Another study on the topic or closely related can be consulted. From the statistical tests or means and standard deviations in the published article, one can often find the likely ES.

Finally, when meta-analyses or individual studies are unavailable, ES can be estimated on a priori grounds (guessing in advance of the study). The investigator may believe that there is no precedent for the type of work he or she is to conduct. (Indeed, this seems to be a fairly common belief among all of us who do research.) The investigator may have to guess whether the ES is likely to be small, medium, or large. The designations are admittedly arbitrary but quite useful guidelines in this regard by noting small, medium, and large ESs to correspond to .2, .5, and .8, respectively (Cohen, 1988). It is helpful to select a conservative estimate. If the investigator is new to an area of research (e.g., first or second study), it is likely that the strength of the experimental manipulation and many sources of variability may be unfamiliar and difficult to control. In such cases, it is likely that the investigator is slightly over optimistic about the ESs he or she expects to achieve and may underestimate the sources of variability that attenuate group differences.

In any case, assume that by one of the above methods we consider the likely ES to be about .50. We have alpha = .05, power = .80, and ES estimated at .50. For actually making the calculation in designing an experiment, most standard statistical packages (e.g., SPSS, SAS, but others as well) provide power calculators. In addition, the Web provides reliable power calculators (e.g., G*Power 3 is one such program: www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/; there are many others) that allow entry of the critical information (e.g., alpha, ES) and the type of statistical test one is interested in and obtain the needed sample size. To get a better feel for critical issues and decisions, I will illustrate the task and challenge from tables provided from a classic textbook on power. The purpose is to better convey the interrelations among the variables that contribute to power and the choices and trade-offs.

13.3.3: More Information on Relations to Alpha, Effect Size, and Sample Size

Table 13.1 provides a simplified table for comparing two means using an alpha of .05. Select values of sample size, effect size, and power are presented rather than all possible values. Hence some of the estimates (exact number of subjects) are very close approximations. Looking at the table gives a better picture of how power relates to sample size and ES than what one would see if computing power for a particular study using one of the methods of calculating power, mentioned previously. In the table, the column marked *n* is the number of cases per group; across the top of the table is ES or *d*, with each column representing a different ES. The entries within the body of the table itself reflect power. So let us enter the table in the column with ES = .50 (circled) just to get an idea of how the power "works." As we go down the column, we are looking for .80 (also circled), which is the minimum level of power we would like for our study. As you recall, this mean an 80% chance of detecting an effect in our experiment if there truly is an effect. The table is marked to show .80, and then the horizontal line moving to the left shows the *n* we need. When alpha = .05, ES = .50, and desired power is .80, we

need 65 subjects per group (i.e., N = 130 for our two-group study). (It is important to underscore that the table gives *n* or the size of each group in the study. N is a different number and reflects the total sample. Stated another way, in a given study there are *k* number of groups. In a two-group study, *k* = 2. Each group will have *n* number of subjects. Thus, the total sample is N = *kn*.)

Table 13.1: Highly Abbreviated Power Table: Power for a Two-Group Study and Testing for Significance (*t* test using $p < .05$)

Group Size (*n*) ↓	Effect Sizes (*d*)					
	.20	.40	.50	.70	.80	1.00
10	.07	.13	.18	.31	.39	.56
20	.09	.23	.33	.58	.69	.87
30	.12	.33	**.47**	.76	.86	.97
35	.14	.38	.54	.82	.91	.98
40	.14	.42	.60	.87	.94	.99
45	.16	.47	.65	.91	.97	.99
50	.17	.50	.70	.93	.98	.99
55	.18	.55	.74	.96	.99	.99
60	.19	.58	.77	.97	.99	.99
65	.20	.61	**.80**	.98	.99	.99
70	.22	.65	.84	.99	.99	.99

Where:

Group size *(n)* = the sample within each group and is not to be confused with N, the overall sample for the entire study, combining all groups

Effect size *(d)* = Cohen's *d*, which is among the most commonly used among the many available effect size measures

Power = the italicized numbers within the table and refers to the likelihood of detecting a difference (rejecting the null hypothesis) if in fact there is one (i.e., the null hypothesis is incorrect). Power of .80 means that if there is really a difference in the population (real world), that would be statistically significant 80% of the times if the experiment were run an infinite number of times.

Notes:

Effect sizes can range from 0 through no upper limit in principle and are continuous. I have included a few values in even numbers (e.g., .20, .40), but the numbers are continuous between all values in the table. The table includes small, medium, and large effect sizes as .2, .5, and .8, which are recognized to be arbitrary but still serve as a reference point in discussing findings.

Power estimates are approximate because I have not used the continuous numbers in group sample sizes and in effect sizes. These differences are very slight but worth noting. For example, power in the table would vary slightly if the group sizes were 30, 31, and 32. Power cannot exceed .99 in the table but gets higher with further decimals. Power cannot be 1.00; a probability of 1 is certainty and is not possible in empirical research.

Many studies in clinical psychology, but other areas of psychology as well, do not have a sample size this large (N = 130), so maybe we can loosen up a bit.

> In fact, after seeing the total sample N we need, we might say, "Who cares about power anyway?"

What if we lighten up and reduce power to .50. That would mean that we have a 50% chance of detecting a statistically significant difference if one were to exist. In the table, power close to .50 is .47 and is circled in the table, so let us look at that. We move to the left of that number and see that we only need a sample size of 30 per group (or N = 60), again referring to Table 13.1. Of course if we are running subjects (e.g., rather than using some online method of running subjects such as MTurk), obtaining a total of 60 obviously will be so much easier than obtaining a total of 130. Yet, the "price" (low power) is high.

Relaxing power in this way is very risky. In my own research, I care a lot about power. I am not going to do many studies in my lifetime, and I am very unlikely to ever do any particular study over again in the same way, so I am not too keen on handicapping myself with weak power for those few studies that I do conduct. So designing a powerful test really is important to get the most sensitive test feasible. If one adheres to the tradition of statistical significance testing, power and its related concepts are absolutely critical and cannot be neglected.

When we consider power in advance of a study, we are likely to learn that to detect such a reasonable (medium) ES, we need a much larger N than we planned or perhaps even than we can obtain. This is excellent to identify in advance of conducting the study. We may then decide to vary alpha (e.g., $p < .10$), to reduce power slightly (e.g., power = .75), or to select experimental conditions (or variations of the manipulation) that are likely to yield larger ESs. Such informed deliberations and decisions are praised when they are completed *prior* to an investigation is completed. Making decisions after the study results are in (e.g., changing one's view of *p* levels, looking at analyses using only part of the data) gives the appearance and may actually be trying to find something statistically significant.

I have commented briefly on the relation of alpha, power, ES, and N and how changing one of these has implications for the others. More about this can be obtained on the Web (by including these as search terms). Also, many statistical software packages allow one to explore how change in one parameter alters others.

The frustrations of methodologists advocating attention to power stem from two sources we have discussed:

1. The long tradition of most studies being underpowered to detect the likely effects their manipulations will show. Weak power is one of the consistencies of psychological research from the 1960s, when Cohen first noted this until the present (Bakker et al., 2012).
2. The ease of calculating power given statistical software, and Web available calculators make the process very simple and allow one to enter any parameter (e.g., N, ES, alpha, power) and to see any or all other parameters.

In light of these considerations, one is surprised to learn from an evaluation of 12 journals in psychology and education over a 5-year period that less than 2% of the studies conducted (or at least reported) power analysis in advance (Peng, Long, & Abaci, 2012). Actually one should not be surprised because we know from psychological and interactions with our partners and children that telling people to do something, even if in their best interest (e.g., stop smoking, cut down on the lard omelets, calculate power so you know what the sample size ought to be) is not a very effective way of changing behavior. Estimating and increasing power before you begin a study fall nicely in this category of apparently ineffective appeals.

One further point about sample size and power is worth noting. Power pertains to the statistical comparisons the investigator will make, including primary and secondary analyses that may divide groups into various subgroups.

For example, the investigator may have N = 100 subjects in two groups. The main comparison of interest may contrast group 1 (n = 50) with group 2 (n = 50). The investigator may plan several analyses that further divide the sample, for example:

- By sex (males vs. females)
- Age (younger vs. older)
- Intelligence (median IQ split)
- Some other variable

Such comparisons divide the groups into smaller units (or subgroups). Instead of groups with ns = 50, the subgroups are much smaller and power is commensurately reduced. One cannot always anticipate the comparisons one will make in the data analyses. The obtained results may call for further (secondary) analyses to help clarify what was found. Even so, the lesson is simple. Ensure adequate power for the comparisons of primary interest long before the first subject is run.

13.3.4: Variability in the Data

Power is a function of alpha, N, and ES.

However, there is more to power than the formula for its computation. Noted already was the notion that excessive variability within an experiment can threaten data-evaluation validity. Variability is inherent in the nature of subject performance in any investigation. However, the investigator inadvertently can increase variability in ways that will reduce the obtained ES.

Let us say for the moment that the mean difference between groups on some measure = 8, i.e., 8-point difference on some scale, questionnaire, or other measure. That is the numerator of the equation for ES.

ES will increase or decrease depending on the size of the standard deviation by which that difference is divided (the denominator). The standard deviation can be larger as a function of the heterogeneity of the subjects (e.g., in age, background, sex, education, and other variables).

The effects of the intervention or experimental manipulation are likely to be less consistent across subjects whose differences (heterogeneity) are relatively great. The heterogeneity of the subjects is reflected in larger within-group variability. This variability, referred to as *error variance*, is directly related to ES and statistical significance. For a given difference between groups on the dependent measure, the larger the error variance, the less likely the results will be statistically significant. As discussed previously, error variance can be increased by sloppiness and lack of care in how the experiment is conducted, by using heterogeneous and diverse subjects who vary on characteristics related to the outcome, and by using measures that have poor reliability. Procedures and practices that reduce or minimize extraneous variability increase the obtained ES and power.

13.4: Ways to Increase Power

13.4 Determine the different ways of increasing power in statistical tests

There are many ways to increase power. These are highlighted earlier; first of course is increasing sample size. This is important because it is the most obvious and the first line of attack. When undergraduate college student samples can be run and a large subject pool is available, that alternative may be feasible and quite useful. Also if the study can be run online (e.g., via MTurk, Qualtrics), a large number of subjects can be obtained in a short time frame (e.g., days). In clinical settings or other settings where patient populations are of interest or assessment is labor-intensive or expensive (e.g., different types of brain scanning), increasing sample size is not always that easy or feasible. Sometimes increasing the sample size is not possible because there are relatively few clients available with the characteristics of interest (e.g., children with a particular chronic disease, cohabiting adults of different

ethnicity raising twins, professors with social skills). Obtaining large numbers of cases might require sampling across a wide geographical area or continuing the study over a protracted period and in fact would preclude conducting the research. In short, increasing sample size is not always feasible, especially if an investigator wishes to complete the study in his or her lifetime. Alas, there are many options and selecting one or more of these can help enormously.

Ways to Increase Power:

1. Increase sample size (N or n/group)
2. Increase expected differences by contrasting conditions that are more likely to vary (stronger manipulations, sharper contrasts)
3. Use pretests/repeated measures that reduce the error term in the effect size

Effect Size Formula

Without repeated measures	With repeated measures
$ES = \frac{m_1 - m_2}{S}$	$ES = \frac{m_1 - m_2}{S\sqrt{1 - r^2}}$

4. Vary alpha (a priori) if the case can be made to do so if, for example:
   - **a.** Classification of groups (e.g., case-control study) is imperfect
   - **b.** Measures are not well established (dubious psychometric properties)
   - **c.** Small effects/differences (ES and significance) are predicted
   - **d.** Consequences of the decision vary markedly as a function of the direction and hence we wish to detect difference in one direction rather than another (e.g., one-tailed and lenient alpha)
5. Use directional tests for significance testing
6. Decrease variability (error) in the study as possible by:
   - **a.** Holding constant versus controlling sources of variability
   - **b.** Analyzing the data to extract systematic sources of variance from the error term

13.4.1: Increasing Expected Differences between Groups

Assume for a moment that an investigator is comparing two different groups (conditions). These conditions are selected to test a particular hypothesis.

In relation to power, the investigator can ask:

- Is this the strongest test or comparison of the different conditions?
- Can the hypothesis be tested by making the manipulation stronger or by establishing a sharper contrast between conditions?

For example, instead of comparing a little versus a lot (of some variable, manipulation, or experience), the contrast would be sharper if one compared none versus a lot or very little versus quite a lot. Another way to illustrate this is to say the investigator is interested in comparing three groups—low, medium, and high levels of depression on some other measures. A stronger test might be to compare the two groups (low and high). The groups are more extreme and more likely to show a stronger effect, assuming that the characteristic of interest (depression) operates on a linear way so that more is worse on some other measures. Also, for a given number of subjects—say 100 in our hypothetical study—a test of two groups each with $n = 50$ subjects is more powerful than a test of three groups, each with $n = 33$ subjects. This is a separate lesson from the main point, namely, for a given sample size (e.g., N = 100), power can be increased by deploying them into fewer groups.

The overall point is that conditions selected in a study might be made more extreme to provide a test of a hypothesis that is likely to lead to stronger effects. As investigators, we ought to ask ourselves at the design stage of the study, "Can the study be designed so that the anticipated ES is large or at least larger than what we were thinking in our first idea for the study?"

What do you think?

For a given sample size and alpha, a larger ES is of course much more easy to detect as a statistically significant effect. That is, power increases if ES is increased. Increasing ES is in part a function of the specific conditions selected and tested in a study.

13.4.2: Use of Pretests

Noted previously were experimental designs that used pretests. From a design standpoint, advantages of using pretests were manifold and included issues related to the information they provide (e.g., about magnitude of change, number of persons who change, and others). The statistical advantages of a pretest are the most universal basis for using such designs. The advantage of the pretest is that with various analyses, the error term in evaluating ES is reduced. With repeated assessment of the subjects (pre- and posttest), the within-group (subject) variance can be taken into account to reduce the error term. Consider the impact on the ES formula we have been using.

As noted earlier, the formula for effects is: $ES = (m_1 - m_2)/s$. When there is a pretest measure or another measure that is related to performance at posttest (e.g., covariate), the ES error term is altered. The formula is represented by $ES = (m_1 - m_2)/s\sqrt{1 - r^2}$ where r equals the correlation

between the pretest (or other variable) and posttest. As the correlation between the pre- and posttest increases, the error term (denominator) is reduced and hence power of the analysis increases. Several statistical analyses take advantage of the use of a pretest, such as analyses of covariance, repeated measures analyses of variance, and gain scores (e.g., Cook & Steiner, 2010). As feasible, using designs with repeated measures is advantageous because of the benefits to power.

13.4.3: Varying Alpha Levels within an Investigation

Alpha levels are quite related to power and hence their use and variation warrant attention. Alpha at $p < .05$ or $< .01$ is rather fixed within the science and represents constraints over which the investigator would seemingly have little control. Yet, there are circumstances in which we may wish to reconsider the alpha level. The investigator may decide to relax the alpha level (reduce the probability of Type II error) based on substantive or design issues that are decided in advance of data collection. By reducing the probably of Type II error (saying there are no differences when there really are), we increase the likelihood of a Type I error (saying there is a difference when there really is none). Yet, in any given circumstance, there may be great reasons to tinker with these levels and where one kind of error is less important than another.

Several circumstances may lead the investigator to anticipate specific constraints that will reduce the likely ES and the differences between groups or conditions:

1. The criterion for selecting groups in a case-control study might be known to be imperfect or somewhat tenuous. Thus, some persons in one group (e.g., non-depressed controls) might through imperfect classification belong in the other group (e.g., depressed persons). Indeed, if a psychiatric diagnosis is required for the depression and not meeting criteria for a diagnosis is required for the control group that is not much help. Just missing the diagnosis (sometimes referred to as subclinical) is very close to depression because the cut point is arbitrary or at least not currently defensible. Comparison of groups will be obscured by variability and imperfect classification. That imperfect classification is analogous to diffusion of treatment as a threat to internal validity. Some non-depressed and depressed clients made it into the opposite group into which they belong.
2. The measures in the area of research may not be very well established. The unreliability of the measure may introduce variability into the situation that will affect the sensitivity of the experimental test. The predicted relation may have been evident with more sensitive and reliable measures. It may be reasonable to be less stringent (e.g., $p < .10$) in identifying effects that are considered to be veridical.
3. The specific comparison of interest may be expected to generate a very small difference between groups. If we expect small differences, the usual advice would be to increase sample size so that power will be high for this small effect. Yet, this is not always possible. Also, a small difference might be quite important theoretically. We would want to detect that.
4. We might alter alpha level based as well on consideration of the consequences of our decisions. Consequences here may refer to patient care (benefit, suffering, adverse side effects), cost, policy issues (e.g., ease of dissemination, providing the greatest care to the greatest number), and other considerations where the weight of accepting or rejecting the null hypothesis has greatly different implications and value. For example, if we are studying whether a particular procedure has side effects (e.g., illness, death), we might want to alter alpha to, say, $p < .20$. In such a study, we may wish to err (Type II) on the side of stating the side effects exist if there is any reasonable suggestion that they do. Indeed, we would not want to say that more people in the treatment died, but this was not a statistically significant difference (e.g., $p < .08$). In this case, we want to ease up on the criterion (alpha) for making that claim. The principle here—slavish adherence to an arbitrary .05 ought to give way to sound reasoning and caution.

More Information on Varying Alpha Levels

In a given experiment, alpha is one of many decision points. Even though the acceptable level of alpha is deeply ingrained by tradition, the investigator ought to consider thoughtful departures based on circumstances of the particular experiment. There are circumstances when the investigator may plan on using different levels of alpha within an experiment. For example, suppose we are studying comparing two ways of manipulating emotion regulation for individuals we expose to some stressful experience in the study. We have three conditions:

1. One strategy to induce regulation
2. Another strategy
3. A control condition where no effort is made to induce regulation

We sample 75 persons who meet various criteria (e.g., no current psychiatric diagnosis, ages 18–25) and assign them

randomly to conditions, with the restriction that an equal number will appear in each group.

What shall we use for our alpha level?

> We could use an alpha of .05 and let the matter rest. Alternatively, we might in advance of the study consider the comparisons of interest and their likely ESs. The difference between the two emotion regulation strategies versus the control condition is likely to be fairly large, which we may know from prior research. The usual alpha level ($p < .05$) to detect a difference might well be reasonable here. In contrast, the difference between the two regulation strategies is likely to be smaller. A sample of 75 subjects with 25 cases per group in our hypothetical study may be way too small to show statistically significant differences. It might be reasonable to use a more lenient alpha level (e.g., $p < .20$) for comparisons of the two emotion regulation strategies.

In general, in a given instance it may be useful to reconsider alpha level before a study either for the entire study or for some of the tests or comparisons. If on a priori grounds special conditions within the design can be expected to reduce ESs, a more lenient alpha may be justified. Both theoretical and applied concerns might lead to reconsidering alpha. Altering alpha level might be guided by evaluating the nature of the consequences of different decisions, i.e., concluding that there is or is no reliable difference between conditions.

> Tinkering with alpha levels has to be considered very carefully. Obviously, relaxing alpha levels after the fact or when the results just miss conventional significance levels is inappropriate and violates the model on which significance testing is based.

It is tempting to relax alpha levels in this way because few believe that a finding has been supported at $p < .05$ but is unsupported at a p level above that (e.g., $p < .06$ or $< .10$). However, within the conventional model of significance testing, some generally agreed-upon criterion has to be selected. Whatever that criterion is, there would always be instances that just miss and in which the investigator, but not many others of the scientific community, would say that the effect is close enough to be regarded as reliable.

Significance tests and their binary decision-making feature (yes, reject null hypothesis; no, do not) raise the problem of alpha here. All to the points still apply if one is using statistical significance as a criterion. However, in virtually all cases, it will be important also to report ES or some measure of the magnitude of the effect or strength of the relation, as discussed below. Statistical significance is too "iffy" and depends on sample size too heavily. ES gives a measure of the strength of the relation and is more revealing and informative. More on that later.

13.4.4: Using Directional Tests

Variation of alpha levels raises a related solution to increase power, namely, the matter of using one- versus two-tailed tests of significance. Consider a two-group study and a t test to evaluate group differences. The null hypothesis is that the groups do not differ, i.e., the ES = 0. A two-tailed test evaluates the obtained difference in light of departures from 0 in either direction, i.e., whether one group is better or worse than another. The alpha of .05 refers to both "tails" and ends of the normal distribution, which are used as the critical region for rejection.

In much research, the investigator may have a view about the direction of the differences. He or she may not wish to test if the ES is different from zero but rather whether the treatment (e.g., mindfulness) is better than the control condition (just talking about every day topics).

> The hypothesis to reject is not bidirectional (better or worse) but unidirectional (better). As such, the investigator may wish to use a one-tailed test. A lower t value is required for the rejection of the null hypothesis if a one-tailed directional test is provided.
>
> Many hypotheses in research are directional in the sense that investigators have an idea and interest in differences in a particular direction. For this reason, long ago some methodologists have suggested that most significance testing should be based on one-tailed tests (e.g., Mohr, 1990).
>
> However, there is resistance to this to which the reader should be alerted. There often is an implicit assumption that investigators who use one-tailed tests may have done so because the results would otherwise not be statistically significant. Often it is unclear to the reader of the research report that the use of one-tailed tests was decided in advance of seeing the results. The implicit assumption does not give the benefit of doubt to the investigator. At the same time, relatively fewer studies in clinical psychology and related areas utilize one-tailed tests. One rarely sees such tests or sees them in situations where the results would be significant whether the tests were completed as one- or two-tailed tests.

In general, investigators are encouraged to be conservative in their analyses of the data and in drawing conclusions about relations that are reliable or statistically significant. Yet, directional hypotheses and use of one-tailed tests warrant consideration and more active use. We are not necessarily better off as a science because of conservatism in using two-tailed tests almost all of the time. Indeed, the higher priority is providing good tests of informed predictions. A directional test of a directional prediction is probably the best match.

When using one-tailed tests, clarify the basis of this use. That can be accomplished by answering such questions as, "Why is the prediction directional?" and "Is there

a compelling reason not to care about effects in the other tail (direction) even if they were large?" In short, why such tests are to be used should be embedded in the goals of the study and described in the introduction to the study and the rationale for the hypotheses (see Ruxton & Neuhäuser, 2010). Explicit statements will help those who conduct studies and those who read the results identify whether the tests are reasonable.

13.4.5: Decreasing Variability (Error) in the Study

The final method of increasing power is decreasing variability in the study. This topic has been discussed previously, and I can be brief as a result. Noted previously has been the fact that increasing variability (e.g., differences between subjects) can stem from many sources, including how heterogeneous the samples are (e.g., children and adults vs. just adults) and how careful the experiment is executed and conducted (e.g., loose protocol effect, monitoring of treatment integrity). Careful control of variability can be achieved by monitoring many facets of the study; many sources of variability can be held constant. The silent rewards of careful work are reflected in minimal variation that otherwise would be counted as error. This translates to a stronger ES in the study that would have been obtained if less care were taken.

Error variability includes all sorts of influences in a study. In fact, somewhat loosely, consider that an investigator is studying a disorder (e.g., bipolar) and compares two groups on several measures expected to relate to the disorder (e.g., attachment patterns, cognitions, response to stress). In this study, we could say that the investigator is studying the variation (differences) that is a function of one variable, namely, the disorder. There are of course many other variables that the investigator is purposely ignoring. For example, not everyone in the study is of the same sex, gender identity, socioeconomic level, ethnicity, culture, and so on across a large number of variables. The influences of all of these virtually infinite other variables form part of the error term (standard deviation, variance) of the study. One can reduce the error variance of a study, by analyzing the data by one or more of the variables embedded in the error variance. Of course, one selects variables that are worth evaluating because they are used to test, explore, or generate hypotheses of interest. So, for example, one might analyze sex differences, if the case could be made that this is worth testing or exploring. The analyses would not just include those with and without the disorder (a *t* test comparing two groups), but also rather disorder by sex analyses (an *F* test in which anxiety, sex, and anxiety *x* sex are terms in the data analysis output). Sex and the interaction of sex with anxiety disorder now become variables that are evaluated separately and reduce sources of variation that otherwise would be in the error term. In short, either holding constant variables that may increase error variation in the study or analyzing variables that might be included in an error term can be used to decrease error variation. One cannot just analyze all sources of error (all variables that can be imagined), because the case could not be easily made that they are interesting (e.g., anxiety, sex, show size, height, and so on as independent variables) or important.

13.5: Planning the Data Analyses at the Design Stage

13.5 Express why the design stage of any robust statistical test needs to account for data analysis

Data analyses do not emerge as a discussion or topic of contemplation after the data are in—that is a methodologist's nightmare. (Another methodologist nightmare is a power outage.) Issues related to alpha, power, and anticipated ESs, to mention a few, are critical to ponder as the study is being planned. These are not esoteric issues or merely quantitative nuances. Rather, they will squarely affect the conclusions the investigator can draw and the strength and quality of the design. More concretely as the purpose of the study and design are being formulated, it is useful to write each of the hypotheses and next to each one to outline the tentative data-analytic strategies that will be used.

In light of the specific tests and analyses, one can ask:

- Do I have sufficient power given the likely effect size?
- Can I vary alpha, sample size, or reduce variability in some ways (e.g., homogeneity of the sample, how the study is run) to augment power?
- Can I increase the strength or potency of the independent variable or magnify the effect that will occur by using different groups in the design or by contrasting conditions (experimental and control) that are likely to produce stronger effect sizes?
- Do I need each of the groups in this study, or can I deploy all of the subjects to fewer groups (thereby increasing power)?
- Will there be other tests related to this hypothesis that will divide the groups further (e.g., contrasting males vs. females) and thereby reduce power?

Addressing and, to the extent possible, resolving these questions at the design stage are very helpful. After the experiment is completed, no doubt other questions and data-analytic issues will emerge and hence not all of the results and plans for their evaluation can be anticipated or even necessarily rigidly followed. At the same time, the

plan for the major analyses ought to be worked out at the design stage so that changes can be made in the design to enhance the facets of data evaluation validity.

13.6: Objections to Statistical Significance Testing

13.6 Report the arguments against the use of the null hypothesis significance testing in statistical analysis

We will soon turn to practices to aid the use and interpretation of statistical significance testing. Yet, before that it is important to convey enduring concerns about null hypothesis significance testing or NHST. The concerns serve as the basis for all sorts of other practices I will mention as ways to present the data and to rely less heavily on statistical significance testing, even within the tradition of NHST.

Although NHST remains the dominant model in scientific research, there is a long-standing view that null hypothesis testing, as currently practiced and interpreted, is misleading, counterproductive, and simply flawed (e.g., Greenland, 2011; Krueger, 2001; Zilliak & McCloskey, 2008). Recommendations have included a range of options:

- Abandon the practice entirely
- Supplement significance testing with other information
- Replace such testing with alternative statistics

Efforts to eliminate and replace statistical significance testing are not new, nor is it a radical minority view held by extremists. In fact, the view has been voiced for decades in psychology with textbooks, series of articles, and task force reports (e.g., Morrison & Henkel, 1970; Rogers, 2010; Shrout, 1997; Wilkinson & Task Force on Statistical Inference, 1999). Also, this is not just in psychology, but also voiced in our sister social sciences (e.g., sociology, political science, economics), biological, and natural sciences (e.g., Kraemer, 2010; Murphy et al., 2009). These lamentations are associated with repeated calls for us to change how we propose and evaluate hypotheses and test results and perhaps even abandoning statistical testing all together. It is important to be familiar with the concerns because they can influence decisions about how to evaluate one's own data.

A brief historical comment is in order. When statistical tests and null hypothesis testing first emerged (Fisher, 1925; Neyman & Pearson, 1928), objections followed challenging the logic and utility of such an approach (Berkson, 1938). From that time, there has been a continuous "crescendo of challenges" (Kirk, 1996, p. 747). Leading researchers in psychology have made strong statements to convey the point. For among the stronger statements (Meehl 1978, p. 817) noted that significance testing to evaluate the null hypothesis, "is a terrible mistake, is basically unsound, poor scientific strategy and one of the worst things that ever happened in the history of psychology." (Of course, as a clinical psychologist, this statement is hard to interpret, but it sounds negative to me.) The objections to significance testing pertain to what they do and do not accomplish and how they are misinterpreted. The objections are mentioned briefly as a way of moving toward an alternative recommendation for statistical evaluation of research. Table 13.2 provides a convenient summary of major concerns about the use of statistical tests and misconceptions surrounding their interpretation. It is useful to consider each of these in more detail.

Table 13.2: Statistical Significance: Common Concerns and Misconceptions

Statistical Significance	Description
Concerns	• All-or-none decision making, • H_o is rarely or never true, • Significance is a function (and measure) of N, • Tests are more subjective than meets the eye, and • Says nothing about the strength or importance of the effects
Misconceptions (i.e., It is not true that. . .)	• *p* reflects the likelihood or the degree to which the null hypothesis (H_o) is true, • Higher *p* value ($p < .0001$) is a more potent or stronger effect, • Higher *p* value is an effect more likely to be replicated, • No difference means that there is no real effect but a difference means that there is, and • There are nonsignificant "trends" or a difference that "approached" significance or is of "borderline significance" ($p < .10$).

13.6.1: Major Concerns

Many statisticians and methodologists and just plain scientists object to NHST in principle and practice:

1. A difficulty with statistical tests is that in their current use they require us to make a binary decision (accept, reject) for the null hypothesis. We set a level of alpha (e.g., $p < .05$) and decide whether or not to reject the null hypothesis, i.e., there is no difference. The criterion of $p < .05$ is recognized to be arbitrary. Perhaps due to training and use of this criterion, researchers tend to believe that findings at $p < .05$ or lower are genuine effects (i.e., there are group differences) and reflect a relation but that above this level ($p > .05$) group differences do not exist or are just chance. There is no rational basis for this. A quote from prominent methodologists that became classic soon after it appeared

conveys this more dramatically: "Surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277).

2. A concern with significance testing is that the null hypothesis is always (or almost always) false. That is, the means for two groups will always be different (e.g., at some decimal) and asking if groups are different "is foolish" (Tukey, 1991, p. 100). This means that whether or not a difference between groups is significant is largely a matter of sample size.

   With a very large sample size, any difference will become significant. Psychology experiments have become fixed at 10–50 subjects per group in much of the research.

   This fixes the science at examining what empirical relationships hold given this sample size—does anyone really care about that in principle? Also, making a binary decision with this sample size is likely to detect as significant only large ESs. Power to detect differences is weak for most psychology studies for small-to-medium ESs. Weak power is exacerbated when we perform adjustments on the analyses to control for error rates that make alpha more stringent for individual comparisons. Thus, there will be a large tendency to commit Type II errors (not being able to detect a difference when there really is one).

3. There is misplaced faith in the objectivity that statistical analyses provide. Yes, it is the case that many facets of statistical tests provide objective criterion for decision making and this is the overriding advantage. That is, at least many of the rules are clear about what is and is not counted as a "real" effect. At the same time, noted previously was the point that there is considerable subjectivity in selection, use, and reporting of statistical tests and in the conclusions that reached from these. In addition, underlying applications of many statistical tests (e.g., default criteria, rules for decision making) are arbitrary and can be changed by the investigator (e.g., when to make cutoffs for allowing variables to enter into a regression equation) and findings can be significant or not significant as a result. Yet, these points apply to statistical analyses in general, and the specific concerns that are most commonly voiced are about using statistical tests in relation to the null hypothesis, i.e., making a binary decision to reject or accept that hypothesis.

4. A major concern has been that tests of statistical significance do not say anything about the strength (magnitude) or importance of an effect. The degree or magnitude of association is more critical than whether the results are statistically significant, especially, because statistical significance is so dependent on sample size. It is not difficult to identify strong effects (large ES) where group differences were not statistically significant and weak effects where group differences were statistically significant.

   An example is provided below to convey how identical findings and a medium-to-large ES can yield both significant and nonsignificant effects. In general, when groups are truly different, we care less about the actual *p* value and more about whether the differences are large, small, or in-between.

   Also, when we consider multiple independent variables (e.g., prediction studies), we want to know their relative impact in relation to some criterion or outcome. So, for example, for a given outcome, we might look at several predictors (e.g., grandparent diet, one's own diet, exercise, education, IQ, and so on) and how they relate to an outcome in adulthood.

   We want to know about the magnitude of the predictor not just whether they are statistically significant. For example, being a middle child and being subjected to harsh physical punishment in childhood, each predicts delinquency and antisocial behavior in adolescence. Yet, the magnitude of these influences is very different—middle child business not much of an effect and weak ES; corporal punishment much greater impact. A statistically significant effect just says yes or no about whether it makes a difference. We would like to know the magnitude of the relation and how much we ought to be concerned and perhaps even use that to decide what to study further or where to take possible action.

13.6.2: Misinterpretations

Leaving aside objections to NHST, there are a number of concerns about how the tests are interpreted by investigators and by consumers:

1. A set of related misconceptions pertain to the obtained *p* value in a study. Many investigators believe that the *p* value reflects the likelihood that the findings are true, the potency or strength of the effect, and the likelihood that a finding will be replicated. Thus, a *p* of < .002 obtained for a statistical test in a study often is considered to be much better than a *p* of < .05. It is the case that such a *p* value may represent a stronger relationship between the independent and dependent variable (depending on the measure used to evaluate strength such as ES or *r*). However, the *p* value is not a measure of any of these. Rather it refers to the likelihood that the finding would be obtained by chance if a large number of tests were run.

2. No difference (statistically) means that there really is no difference in the world. That is, when we accept the no-difference hypothesis, this means that in fact the effect of the manipulation is zero or that conditions (e.g., experimental vs. control) really are not different. This is

a misconception because no-difference can depend on many aspects of the study, including sample size, but also the conditions under which the hypothesis was tested (e.g., sample, how the manipulation was operationalized), and the care with which the study was conducted (e.g., how much holding constant, tight rather than loose protocols). Also, "chance" too is operating and our finding could be one of those instances in which a real difference was not evident in the sample we selected. There are many circumstances in which a no-difference finding is quite interpretable, but as a general rule it is not quite right to assume no difference on a statistical test means no difference or no impact of one variable on another in fact.

3. Statistical significance is often interpreted as substantively important or significant in some other way than statistical significance. In everyday parlance "significance" is close to the meanings of "importance," "consequential," and "meaningful." Naturally, we, as investigators, move with regrettable ease in noting that a statistically significant effect is significant (meaning important). The change in use or meaning within the write-up of a study usually can be detected in the Results and Discussion sections. The Results focus on the statistical significance, but the Discussion may move to noting the findings are significant, meaning important. Am I making this up? In one review, over 800 articles across health science (e.g., epidemiology, medicine) were evaluated spanning a 10-year period. In total, 81% of the articles were identified as having misinterpreted statistical significance as being significant in a substantive way (i.e., importance) (Silva-Ayçaguer, Suárez-Gil, & Fernández-Somoano, 2010). The term "significance fallacy" is sometimes used to refer to the interpretation of statistical significance as being a measure of "real" significance or importance.
4. For many of us as authors, null hypothesis statistical testing requires regrettable and adherence to rigid p levels to make decisions. The common use of $p < .05$ or $< .01$ reflects the threshold we must use for making a binary decision of accepting or rejecting the null hypothesis, as said before. Strictly, the tests are conducted to see whether the criterion of p is met or surpassed. Thus, one ought not to discuss "trends" in the data (a trend is a slope so that the term is not properly used anyway). So, if the statistical test shows that $p = .07$, one does not say the test "approached significance." Rather, the test is not significant! There is not a "trend," "almost significant effect," or "borderline significance or effect," or other wildly creative terms. (Again, this is precisely why I adopted as alpha $p < .33$ for all my tests—my advisors hated it, but it solved a lot of problems for me.)

13.6.3: More Information on Misinterpretations

By the rules of NHST to which most scientists implicitly agreed, a nonsignificant finding is just that. If one wishes to detect almost-significant findings, then there are all sorts of options such as those discussed previously to increase power. Other options are available as well such as emphasizing magnitude of effect rather than NHST. However, within the rules of the science and null hypothesis testing game, nonsignificance ought to be adhered to in deciding whether to reject or accept the null hypothesis. Terms, such as "approached significance," that authors often use have no formal meaning in science and refer loosely to "almost" rejecting the null hypothesis. Near misses are not what this is all about. They count as misses.

Mentioned above was the fact that a higher level p value (e.g., $p < .0001$ rather than $p < .05$) does not necessarily mean that the effect is more potent, stronger, or important or more likely to be replicated in a subsequent study. Authors often use terms that belie a misconception about this. One can identify authors referring to "highly significant effect." This usually means that p was much less than the alpha selected. The term is amusing because there is an implication that the finding may reflect highly significant in some other way than statistical, and it may not. Bottom line: as you read research studies, smile if you see any of these in the manuscript: "showed a trend toward significance," "approached significance," "was significant at $p < .10$," and "was almost significant." Statisticians and methodologists have a whole series of jokes beginning with, "A guy walks into a bar and says, 'approached significance.'" Be sure to avoid using them in your own work so that your colleagues will not smile. In all of this business of terminology, I am not advocating the general approach of null hypotheses statistical testing. Rather I am conveying that once one does adopt that approach, squirmy terms and thinking (e.g., "my effects almost approached borderline significance") are not appropriate and misinterpret how NHST works.

13.6.4: Significance Testing and Failures to Replicate

Significance testing may impede replication and the accumulation of knowledge. There are all sorts of contradictory findings and failures to replicate. To be sure, many of these might come from the fact that a given finding may depend on moderators (e.g., age, sex, social class of the population) and variations in the samples among the different studies and in any given case one cannot rule out chance, i.e., the effect was one of those 5% of the studies that showed the effect (using $p < .05$) on a chance basis if the study were repeated a large (or infinite) number of times.

> In relation to the present discussion, we are confronted with a more dramatic point, namely, that identical findings can yield contradictory results and conclusions, when statistical significance testing is the basis for drawing influences.

Consider for a moment that we have completed a study and obtained an ES of .70. This magnitude of effect is one that is about the level of ES demonstrated when psychotherapy is compared with no treatment. An ES of this magnitude indicates a fairly strong relation and would be considered as a moderate-to-large ES. Would an ES of this magnitude also be reflected in statistically significant group differences? The answer depends on the sample size.

Consider two hypothetical studies, both with an ES of .70:

- In Study I, we have a two-group study with 10 cases in each group (N = 20).
- In Study II, suppose we have two groups with 30 cases in each group (N = 60).

(Although this example is hypothetical, there are scores and scores of studies with sample sizes in the range of both examples.) We complete each study and are ready to analyze the data. In each study, we have two groups, so we decide to evaluate group differences using a *t* test. The *t*-test formula can be expressed in many ways. The relation between statistical significance and ES for our two-group study can be seen in the formula:

$$t = \text{ES} \times \frac{1}{\sqrt{1/n_1 + 1/n_2}}$$

where ES = $(m_1 - m_2)/s$.

In Study I, when ES = .70 and there are 10 cases in each of the two groups, the above formula yields a $t = 1.56$ with degrees of freedom (*df*) of 18 (or, $n_1 + n_2 - 2$). If we consult a table for the Student's *t* distribution, easily obtained from many sites on the Web we note that a *t* of 2.10 is required for $p = .05$. Our *t* does *not* meet the $p < .05$ level, and we conclude no difference was found between groups l and 2.

In Study II, when ES = .70 and there are 30 cases in each of the two groups, the above formula yields a $t = 2.71$, with a *df* of 58. If we consult Student's *t* distribution, we note that the *t* we obtained is higher than the *t* of 2.00 required for this *df* at $p < .05$. Thus, we conclude that groups 1 and 2 *are* different. Obviously, we have two studies with identical effects but diametrically opposed conclusions about group differences. This is chaos and not how we want our science to proceed. Also, this harkens back to the refrain that statistically significant differences primarily are a measure of sample size.

In this example, identical results yielded different conclusions. The implications on a broad scale are enormous. When we express skepticism in noting that a finding was found in one study but not replicated in another, this is based on the fact that in one study the results were statistically different and in another study they were not. In the accumulation of knowledge, we have not separated those failures to replicate that in fact reflect similar results (ESs) from those that represent genuine differences in the findings.

13.6.5: General Comments

We have discussed the NHST, how it "works," and key objections. While it is important to know these matters generally, the discussion has very practical implications to help with one's research. It is important to reiterate that the dominant model of research continues to be NHST. What that means practically is that it is helpful to be fluent in practices that can optimize the effects of finding differences when they exist.

Obtaining statistical significance from the data analyses in a study is a function of many different features of an experiment, only one of which is whether there is a relation between the independent and dependent variables.

> Testing for statistical significance depends on multiple, interrelated concepts. The researcher ought to know the concepts, how they interrelate, and how to "control" them.

The main reason is that a true effect (difference in the world) may not come out in an experiment (no difference) based entirely on not utilizing information about the core concepts on which statistical significance depends.

13.7: Hypothesis Testing: Illustrating an Alternative

13.7 Examine the Bayesian data analyses as an alternative to the null hypothesis significance testing in statistical analysis

It is important to mention an alternate way of looking at the data than null hypothesis testing.

13.7.1: Bayesian Data Analyses

In this section, I provide an overview of Bayesian data analyses. I mention the approach because it is an alternative to NHST, is gaining use in psychology but other sciences as well, and is important recognize as an option. That said, this text is not about the details of statistical tests and analyses. Bayesian analysis is a family of data-analytic procedures, and various textbooks and courses will be needed to fully appreciate the options. Even so, in the context of NHST and concerns, it is important to recognize that alternatives exist. The purpose of this section is to introduce a

few concepts rather than to foster mastery of a novel way of making predictions and analyzing data.

Bayesian analyses method can be traced back to Thomas Bayes (c 1701–1761), who was an English mathematician and minister. Among his contribution is the Bayes theorem, which connects ones initial (pre-data collection) belief or hypothesis and then looking at the data that were collected to evaluate the likelihood of that hypothesis. That is, we have an alternate hypothesis that something is likely—treatment will be better than no treatment or depressed patients will perform worse on this cognitive task than non-depressed patients. We want to test the likelihood that our hypothesis explains the data.

As a useful point of departure, NHST is based on rejecting the null hypothesis. Yet, in research and life, rarely do we care if the null hypothesis is true. In fact, as mentioned in my earlier quotes in this chapter, one reason we do not care is that the null is never true. But more importantly, we want to know about our hypothesis, the one we want to test and the one we believe to be true. Even if we rejected the null hypothesis in the traditional way, this says nothing about our hypothesis and whether it might account for the data.

> Bayesian analysis pits the null hypothesis against our alternative hypothesis and asks which hypothesis is more credible given the data we have just collected.

Based on the obtained data in a study, the approach evaluates how plausible or probable is the specific theory or hypothesis we are testing. Bayesian methods incorporate our predictions and views about what might happen and then consider the probability of observing an outcome given that our hypothesis is true. Most of the time, we have one or more sources on which to draw (e.g., prior research, expertise, experience, intuition, ideas cleverly drawn from a seemingly unrelated area of research). We have a prediction of what will happen and want to test that.

We begin with our expectation of what will happen. This is converted to a probability (called *prior probability* because it is before actually doing the study). Data are collected and the probability is estimated about the extent to which the hypothesis is true given what the observed data show.

> *Likelihood* is a term used to refer to the probability of obtaining the exact data we obtained given our hypothesis.

The evaluation of the effects comes from the approach by integrating these probability distributions and making a decision about the likelihood of the prior beliefs about the relationship. One has to choose whether the prior belief is likely to be true. This relies on the Bayes factor, or a ratio of the probability of obtaining the observed data given the null hypothesis, divided by the probability of obtaining the observed data under the alternative hypothesis. That ratio provided by the analysis has various cutoffs that can be used to reflect weak, moderate, or strong (and gradations in-between) support. The specific cutoff elected can be influenced by the benefits of deciding correctly and the cost or disadvantages of deciding incorrectly. I have only provided highlights of the approach and only to convey that there are alternatives to NHST.[4]

13.7.2: More Information on Bayesian Data Analyses

The use of Bayesian data analyses has been advocated for decades for use in social and natural sciences and medicine but has not caught on very well. For example, it is easy to thumb through the complete issue of major or minor psychology journals and not find one article using the Bayesian analyses. Also, the analyses are not likely to be taught in statistics courses in most graduate research programs in psychology. Among the reasons is that resources were not as readily available as they are now to make teaching of the techniques and using the analyses very user-friendly. That has changed; introductory text material, methods to calculate critical features required for the analyses, and software programs are now available to allow researchers to draw on and use the approaches (e.g., Jackman, 2009; Kruschke, 2011a). Consequently, many anticipate Bayesian analyses will greatly increase in use in the coming years.

It is important not only to be aware of the use of Bayesian data analyses but also the fact there is controversy and debate about the utility and relative merits of both traditional NHST and Bayesian methods (Gelman, 2008). There are many issues, and they have their rebuttals. For example, selecting prior probabilities about the likelihood that one's hypothesis is true may seem unduly subjective. That point is arguable and can be surmounted in varying degrees (e.g., Dienes, 2011). The debates about Bayesian analyses, in my view, do not argue for one data-analytic approach at demise of another.

> The larger lesson of methodology is that different designs (e.g., randomized controlled trials, single-case), methods of assessments (e.g., brain imaging, computerized assessment), and methods of data analyses (e.g., within traditional statistical significance testing or significance testing vs. Bayesian analyses) can and often do influence the conclusions one reaches. We want to be sure that key findings and principles we hope to demonstrate are not unduly connected to one or two ways in which we evaluate them.

NHST and Bayesian analyses have a different way of approaching hypothesis testing, but that does not mean the results will disagree. An analyses of hundreds of studies found that traditional statistical analyses and Bayesian analyses often are consistent (e.g., when p values are small

and ESs are large, Bayes factors are large) when the experimental effects are strong (Wetzels et al., 2011). The different approaches still yield different information, especially in relation to the plausibility of the alternative (non-null) hypothesis.

In general, science profits from diversity in data-analytic methods and in other facets of methodology, not for some blind ecumenicism that welcomes all approaches for their own sake. Rather, we look for consistencies across studies and approaches and want to understand the differences when they do emerge. As such I would encourage any researcher to learn and test the Bayesian analyses to understand, see, and evaluate the yield. There are excellent materials I have already mentioned, but in addition a YouTube descriptions (www.youtube.com/watch?v=IhlSD-lIQ_Y). Also, there are excellent comparisons of NHST and Bayesian analyses with the same data sets to convey differences and similarities in the approach and yield (e.g., Kruschke, 2013; Wetzels et al., 2011).

13.7.3: General Comments

Null hypothesis testing is not the only way in which statistical evaluation of results can be conducted. At this point in time, we know that there has been a continuous literature lamenting the use of significance tests that focus on binary decisions and recommending alternatives. Indeed, R.A. Fisher, who is credited with (or, given fickle history, blamed for) beginning significance testing, recommended that researchers supplement their significance tests with measures of the strength of the association between independent and dependent variables.

We would like to know more than whether a null hypothesis can be rejected; we want to know about the size or magnitude of the effects we study and whether they are very strong. Efforts have been made to move away from the exclusive focus on statistical significance testing in favor of methods highlighted here. There are alternatives to supplement and complement significant tests.

Summary and Conclusions: Null Hypothesis Significance Testing

Null hypothesis significance testing is the dominant method to analyze the results of research. Statistical tests use probability levels to make the decision to accept or reject the null hypothesis that there is no difference. The decision emphasizes the concern for protecting a Type I error, i.e., rejecting the null hypothesis when that hypothesis actually is true. Because statistical significance remains the primary criterion in evaluating results of research, it is incumbent on the researcher to understand how to design studies that have a strong chance of demonstrating differences when they exist.

Issues critical to statistical evaluation were discussed, including:

- Significance levels
- Power
- Sample size
- Significance and magnitude of effects
- Multiple comparison tests
- Multivariate data

Statistical power has received the greatest discussion in research because it shows most clearly the interrelation of alpha, sample size, and ES. Evaluations of research have shown repeatedly that the majority of studies are designed in such a way as to have weak power. The obvious solution to increasing power is to increase sample size, although usually this is not very feasible, in part because adding too many rather than just a few subjects is often required. Additional strategies to increase power include using stronger manipulations or more sharply contrasted experimental conditions (to increase ES), using pretests or repeated measures to reduce the error term, varying alpha (planned in advance of the data analyses) in selected circumstances outlined previously, using directional tests of significance, and minimizing error variability in all facets of the experiment to the extent possible. There is much an investigator can do, even when sample size cannot be increased.

A critical facet of data analysis occurs in the planning of the study. Not all issues that arise can be anticipated (e.g., intriguing findings that call for additional analyses). Yet several issues can be addressed at that stage. Multiple questions were raised that one can ask oneself at the proposal stage. Answering these at the outset of the study can increase the likelihood of obtaining group differences if they really exist.

There has been ongoing dissatisfaction since statistical significance testing emerged about the utility of this approach for research. Among the many concerns is the

fact that null hypothesis and statistical significance testing give us arbitrary cutoff points to make binary decisions (accept or reject the null hypothesis), and most importantly does not provide the critical information we would like (e.g., direct tests of *our* hypotheses and information about the strengths of our interventions).

Null hypothesis statistical testing is not the only way of approaching the data and data analyses. Bayesian data analyses were highlighted as an alternative to null hypothesis statistical tests. The analyses, actually a family of analyses, focus on the alternative hypothesis, and the likelihood of the hypothesis we wish to test accounts or is plausible from the data we actually obtained. Bayesian analyses are in active use but not routinely evident in psychology journals in clinical psychology. That could change in light of the enduring dissatisfaction with null hypothesis statistical testing and with more and more user-friendly resources (instructions, software) to conduct Bayesian analyses.

The chapter discussed the dominant model of data analyses and key considerations that relate to the design of study (e.g., power, which groups or conditions to include to optimize the effects). It is important to be armed with knowledge of strengths and limitations of null hypothesis statistical testing in part to address them in novel ways (e.g., supplementary analyses, tests, and ways of looking at the data). In the next chapter, we will consider more concrete facets of data analysis and decision-making in what to present in the data and issues that are likely to confront the investigator in analyzing the data.

Critical Thinking Questions

1. What are the key characteristics of null hypothesis statistical testing (NHST)?
2. What is statistical power? Why is it so important? How can it be increased in a study?
3. Statistical significance and what it means is so often misinterpreted. If you read in an article that says a finding was significant at $p < .01$, what exactly does that mean?

Chapter 13 Quiz: Null Hypothesis Significance Testing

Chapter 14
Presenting and Analyzing the Data

Learning Objectives

14.1 Analyze the nuances of data evaluation

14.2 Evaluate other statistical tests to overcome the limitations of statistical significance testing

14.3 Report the presence of major decision points while doing data analysis

14.4 Identify some of the ways to manage issues that arise in statistical analysis when subjects drop out

14.5 Investigate the rationale of deleting the outliers in a statistical experiment

14.6 Examine statistical analyses that involve comparison of multiple subject groups

14.7 Compare multivariate and univariate analyses

14.8 Report some key considerations on decision points in statistical analysis

14.9 Evaluate the meaning of "exploring the data" in statistical analysis

This chapter focuses on practical issues about how to evaluate and present one's results. This information can be used to complement tests of statistical significance and decision making in evaluating and presenting the data. The chapter begins with data evaluation. With all its objections, null hypothesis significance testing (NHST) still dominates and as such the researcher (and reader) ought to be skilled in the approach, mindful of its liabilities, and have an overflowing quiver of options to improve the yield from one's research. In this chapter, we will discuss practical issues but not specific statistical tests and options and what and when to do them.

14.1: Overview of Data Evaluation

14.1 Analyze the nuances of data evaluation

Let us begin by highlighting some of the basics. You have the data in hand, i.e., all the subjects are run and you have their responses on the dependent measures (e.g., scores on questionnaires, behavioral codes where frequency of some behaviors was assessed, numerical responses to automated measures of arousal, and so on). You know the major statistical tests you will use to evaluate the hypotheses, because these were planned before the study began (please say that you did that—thank you). Before we test the hypotheses and the fascinating parts of the evaluation, we have some preliminary tasks. These are not minor but some are clerical.

14.1.1: Checking the Data

To begin, before the analyses, how can we be assured that the data have been correctly coded and entered on the database we will use for the analyses? The answer may well vary with the type of measure (e.g., fMRI, responses on a task presented on a laptop, and so on). Yet, there may be many opportunities for inaccuracies. Paper-and-pencil measures (e.g., self-report questionnaires) are still in active use but more frequently are presented and scored automatically (e.g., Qualtrics), so the numbers go right into a database. Even so, there are errors in the data, and these could be due to subject responses that were accidentally aberrant and problems in accurately coding the data.

The overarching questions:

> Where might inaccuracies enter into the data, and how can I check or recheck them?

For example, if data are entered by hand on a database, be sure that the data are entered twice and as part of

a data-checking program that immediately identifies errors when the second entry of the data is discrepant with something from the first entry. Be as assured as you can that the data are entered accurately.

Once you feel assured, do some very early descriptive analyses and look at each measure. Specifically look at the range (hi, lo numbers) for each measure to be sure that all scores on the measure fall within the appropriate range. There might be a measure where the scores can go from 10 to 100 and you see that one person has a 5 and another 110. Checking on the range does not capture all errors of course, but this is only one way of checking. Look at the distribution of scores (plot the scores as part of these description). Are there scores (individual dots on a graph) that really standout because they are so far away from the mean? You may see a roughly normal distribution in a scatter plot but one score wildly out of the distribution. Just check—Is the score correct, or was there some error in coding or entering the data?

Is the score correct or was there any error?

Again, the ways to check depend on the measure and how the data are collected. But the overall point is the one that serves as a guide. Be as sure as you can that the data on the database are accurate. The larger the assessment battery, the more diverse the measures, and the fewer resources to check (e.g., research assistants, professional database people who enter and check data), the more likely there will be errors.

One checks on the data for obvious and not-so-obvious reasons. For the obvious reason, we want accurate information—that was the whole point of doing the study. We are trying to find out how things are in the world, and accurately obtained, recorded, and entered data are central to that.

> For the not so obvious reason, errors can wreak havoc on the data. Errors (incorrect numbers entered for a subject) can change both the mean of the group the subjects are in and add variability.

For example, a couple of subjects really scored 60 and 85 on two measures, let us say, but with normal errors, typos, and misentry of data are now on the database as 06 and 58, respectively. Of course, these misentries change both the means but also the standard deviation. If they were off by 1 point (61 instead of 60 and 86 instead of 85), the damage would be less on means and measures of variability. Yet the amount of influence or damage is not the entire point—let us check and maybe double-check. Remember the original hypothesis probably sounded like this. I predict that groups will be different as a function of my experimental manipulation. The hypothesis is NOT—I predict . . . even when lots of errors may be in the raw data. Worse errors are possible than one or two number being off—sometimes all of the data are shifted over one column or so on the data base so that every number is wrong.

Data-evaluation validity usually is not discussed from the standpoint of errors. Errors introduce inaccuracies of course, but their influence on the final outcomes of studies is hard to evaluate. No doubt some studies are not published because the findings did not come out and other are published because the findings did come out, and in both cases one must assume errors could have contributed. Preparation of reports and publication of research and grant applications do not ask about data checking. It is assumed. I mention it here because accuracy should not be assumed. We are humans and that means, among many other things, we design as many procedures as well to check on and overcome our limitations. Checking the data in many ways is part of that.

14.1.2: Description and Preliminary Analyses

Ok, the data are checked and double-checked. In fact, some of your peers think *you* should be checked and specifically for obsessive-compulsive disorder. Do not listen to them. You did the methodologically right thing. Now we want to look at the data in a very preliminary way.

> Consider statistical evaluation as including two main parts: describing the sample and drawing inferences about the impact of the intervention.

We begin now with the description.

For the description portion, here we want to describe the sample. Essentially, who are these subjects (age, ethnicity, diagnoses if relevant)? If one is preparing a manuscript, this information is likely to go in a section describing the participants. Also, as part of the description are the groups different at the beginning of the study on variables that may influence the conclusions. If this is a true experiment, groups were randomly assigned to conditions. It is worth a check to see if the groups are different on key variables; random assignment by definition will lead to differences occasionally. Any differences are important to identify and may be variables that are evaluated to see what their impact might be on the dependent variables.

> In describing the sample, usually we want some measure of central tendency (mean, median, mode) and variability (range, standard deviation). If the variable is categorical (sex, presence of a diagnosis), then the percentage within each group is the suitable statistic.

One cannot elaborate all the variables that might be relevant to the study. But the overall thrust is clear; we are doing some analyses to describe the sample but also the check on any group differences. These differences might relate to initial sample differences but also differences that emerge over the course of the study (e.g., who entered the

study but did not complete or have complete data for some reason). It is difficult to convey precisely what descriptive analyses ought to be provided in any given study. That is why it is useful to be guided by the questions: What is this sample like? Are there any differences between groups on key variables that might influence the results?

The measures used in the study may require special description as well. If the measures are familiar standardized measures (e.g., Beck Depression Inventory) or reflect procedures that are well known with established or widely used scoring method (e.g., eye tracking, fMRI), then the description of the procedures and scoring in the method section may be sufficient. Many studies used home-made measures that are developed for the purpose of a particular study. In many such cases, the measure focuses just a few items to measure a critical construct. For example, the investigator may want to measure concern, forgiveness, or empathy and develop three items to get at that. They will then use the total in the data analysis to measure the construct. This is methodologically weak, to put it mildly, because at best the items have face validity. As we have discussed, that is not to be confused with knowing what the items actually measure. Because the measure seems to reflect empathy, for example, has no necessary or empirical connection to the construct that is sampled. Another construct related to empathy might be measured. Or the scores may reflect socially desirable responding, compliance, or some personality characteristic. As an investigator, include some data analyses to describe performance on the measure. To that, include something to address validity of the measure. The question to address: What evidence is there to support the validity of the scale? Sometimes investigators report reliability measure (e.g., internal consistency), which is lovely but not relevant to the point. What is the evidence for validity? Describe that as part of the descriptive material.

Another measurement issue is worth considering at this preliminary stage. You may have used several measures and expect the impact of your experimental manipulation on these measures. Some of the measures may be home-made, some may be standardized, and so on in various combinations. It is a useful exercise to correlate all of the measures with each other.

What do you think are the inter correlations?

The reason is that you might believe that each measure assesses a construct and is distinct. Yet, it may be that two or more measures correlate pretty highly (e.g., $r > .70$). If the correlation is high, it might be useful or wise to combine the measure into one single measure. This can be easily accomplished by converting each measure to a standard score (e.g., with a mean of 50 and a standard deviation of 10) and summing to reach a total. The single measure might be more reliable index of the construct. Also, when you conduct statistical analyses, reducing the number of statistical comparisons is an advantage. Combining measures that in fact overlap can do that.

In some studies, multiple measures might be used to define a construct (e.g., latent variable) and here the statistical analyses may consider the combination. I am referring to all the other studies in which multiple measures are used and the assumption is made that separate measures are not very related to each other and therefore measure separate constructs. I suggest peeking at the correlations to see if that assumption is supported by the data.

In general, investigators usually prefer to retain the individual identity of a measure and not to combine measures. The main reason is that a combined measure may make statistical sense but is otherwise difficult to interpret and integrate with other studies where the measures were evaluated separately. This is a cogent concern. At the very least, look at the correlations. Whether you combine the measures, the correlations may help. You are concluding that some manipulation altered this, that, and more. It would be useful to know that—this and that were highly correlated and arguably are not separate outcomes.

> You have finished the description of the sample and have evaluated the measures to death. Now we are ready to move on to hypothesis testing. Here now is where you have selected among many different types of analyses that are suited to your hypotheses. The specifics here are the core part of courses on statistics and hence beyond the present scope. Yet, what is within the scope is recalling the discussion of NHST.
>
> Whatever the analyses, it is likely that somewhere you are testing for statistical significance. That may be to reject the null hypothesis in a simple comparison of groups (e.g., *t* and *F* tests), to examine the impact of multiple influences (e.g., hierarchical linear regression), or to evaluate a statistical model of how variables are related to each other and some other outcomes (e.g., mediation analyses, structural equation modeling).

The tests of statistical significance raise concerns we have discussed previously, namely, conclusions that are usually binary (statistically significant or not) and that may depend on somewhat arbitrary features of the study (e.g., how many subjects you used).

14.2: Supplements to Tests of Significance

14.2 Evaluate other statistical tests to overcome the limitations of statistical significance testing

We have lamented the limits of statistical significance testing. Yet this is the method of hypothesis testing that

remains dominant. To surmount some of the objections (e.g., binary decisions, differences that depend on sample size because group means are virtually always different), there is much to do to supplement statistical significance. A variety of other statistics can be presented to enhance the clarity of the findings and that move beyond merely noting yes or no—the effects were statistically significant. Here are major contenders to add to the results in any study in which there is testing of hypotheses via statistical significance. Table 14.1 summarizes these, but they are discussed here in further detail.

Table 14.1: Alternatives or Supplements to Significance Tests

Alternatives or Supplements	Description
1. Magnitude or Strength of Effect	Familiar examples include effect size (ES or Cohen's *d*), *r*, r^2, *R*, R^2, but there are many others (omega^2, eta, epsilon^2)
2. Confidence Intervals	Provide range of values and the likelihood that the ES in the population falls within a particular range CIs = m ± $Z_\alpha s_m$ Where m = the mean score; z_α = the *z* score value (two-tailed) under the normal curve, depending on the confidence level (e.g., *z* = 1.96 and 2.58 for *p* = .05 and *p* = .01); and s_m = the standard error of measurement, i.e., the estimate the standard deviation of a sampling distribution of means or the standard deviation divided by the square root of N (s_m = s/□ *N*).
3. Error Bars for Graphical Display of the Data	Error bars refer to a measure of variability usually plus and minus the mean that is actually graphed with the data. CI is one measure of that range of the interval based on 95% or 99% confidence, which reflects *p*s < .05 and < .01, respectively. This upper and lower limit of the bar is computed in the identical fashion as the CIs and is drawn on the graph.
4. Meta-Analysis	Extends the use of ES across many studies—a way to combine studies, to review a literature, and to identify better estimates of population parameters (ES, CIs)

14.2.1: Magnitude and Strength of Effect

In addition to statistical significance testing, it would be helpful to report some measure of the magnitude or strength of the relation between the independent and dependent variable or the magnitude of the differences between groups. Statistical significance gives us a binary (yes, no) decision but beyond that we want to know much more such as the relation of our variables to each other. This is especially important because statistical significance is so dependent on sample size and as we have noted. Add to that the fact that so often studies are underpowered and not likely to detect a difference. All the more, we do not want to rely only on statistical significance. Magnitude and strength of effect give an idea about the connection of the variables we are studying.

There are two broad categories of measures worth noting:

1. Effect size (ES) measures that focus on the *mean differences between groups*. This is the type of measure we have been discussing most in this course to illustrate points about methodology (i.e., various practices such as unreliability of procedures and measures and the impact they exert). Familiar measures of this type are Cohen's *d*, which we have been using, but there are many others (e.g., Hedges *g*, Glass's Δ).
2. Correlational measures that focus primarily on *shared variance or overlap of the variables that are studied.* Familiar measures include *r* and r^2 (or from regression analyses *R* and R^2), but here too there are other measures in the family of correlations such as omega^2 (ω^2), eta^2 (η^2), epsilon^2 (ε^2), and phi (φ) that cover slightly different circumstances. Correlational measures focus on the proportion of shared variance of the two variables (e.g., independent and dependent variable).

Actually, there are many such measures of magnitude and strength of effect and not all fit squarely in one of the two categories (e.g., Ferguson, 2009; Grissom & Kim, 2011; Rosenthal & Rosnow, 2007). Mean difference-based measures or shared variance measures both reflect indices of magnitude and strength of effect. And one can be converted to the other, as I note below.

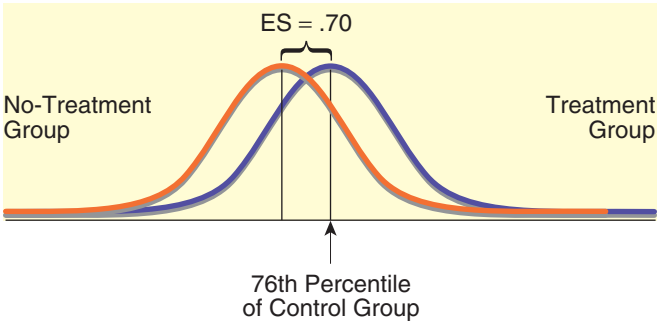
NOTE: The standard error of the mean, noted here as s_m, refers to the estimate of the *standard deviation of a sampling distribution of means*. That is, the mean of a study is an estimate of the mean in the population. If one were to run the study many different times—indeed an infinite number of times, each drawing a random sample of subjects for the population, each study would yield a mean. These means form a sampling distribution of means, i.e., each mean is a data point. The overall mean or the mean of these means would provide the real or population mean μ. But not all the means that were sampled would be the same; they would vary a bit. The standard error of the mean is the standard deviation of the sampling distribution of the means and reflects how much sample means may be expected to depart from the population mean. In a single study we conduct, the standard error of the mean helps us to estimate, with some level of confidence, the likelihood that the population mean will fall within the range we present. If the standard error of the mean is small, then when multiplied by ±

the z score (1.96), the range will be relatively small and we can be reasonably assured that the population mean is within the range.

I have been discussing Cohen's d as a consistent way of discussing methodological and statistical issues. The measure is one of the most commonly used in psychology when ES sizes are presented (Fritz, Morris, & Richler, 2012). The measure of the magnitude of effect based on mean difference is the index most commonly used in individual studies and in meta-analyses that combine the results of several studies. For Cohen's d, the numbers can be interpreted as reflecting standard deviation units. For example, in a study comparing an intervention and control group, a d of .70 is readily interpretable in relation to the differences in the distributions between treatment and no-treatment group. That is, ES can be translated into more concrete terms. Figure 14.1 shows two distributions, one for the treatment group and one for the control group. The means of the group (vertical lines) reflect an ES of .70, i.e., the mean of the intervention group is 7/10 of a standard deviation higher than the control group. One can go to a table of the normal distribution and convert this information into how persons in the intervention group fared relative to control subjects in standard deviation units. Given the ES of .70, the average subject who received treatment is better off than 76% of the persons who did not receive treatment. This percentage was obtained by identifying what percentage of the population is below +.70 standard deviation units on the normal distribution.

Figure 14.1: Representation of an Effect Size of .70 between an Intervention and Control Group

Each group is reflected in its own distribution (normal curve). If the groups in fact are not different, the two distributions would be superimposed on one another and in fact look like one distribution (same mean, same standard deviation).



The most familiar measure for evaluating shared variance is r^2.

The correlation (r) reflects the association or the amount of covariation of two variables (i.e., how much do the independent and dependent variables covary together). The correlation squared (r^2) is used to reflect shared variance in these variables. When analyses of variance are used, eta squared or partial eta squared is the correlation measure usually used (see Richardson, 2011). (When in multivariate analyses of variance or repeated measures analyses of variance measures are used, the observations are not independent in some way and partial eta squared is used.) One can readily see how proportion of shared variance is derived and what this means from the formula by looking at eta squared, which as I noted is commonly used to follow analyses of variance.

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

Where:

SS_{effect} = the sums of squares for the effect of interest (a factor in the design of an ANOVA); SS_{total} = the total sums of squares for all effects, interactions, and errors in the ANOVA. That formula will yield a ratio with a decimal (e.g., .20); remove the decimal and add a percentage sign (e.g., 20%) to have the proportion of shared variance.

Eta and other measures of effect size or proportion of variance (Cohen's d and r^2) are easily computed from formulae provided in introductory statistics textbooks and from Web sites, as well as commonly used statistical software packages. Also, these estimates can be computed directly from familiar statistical tests of significance for comparing two groups. Note the easy conversions in Table 14.2. These convey that once one has a t or χ^2, one can provide further information on ES and r. Thus, in terms of reporting results, one can derive with relative ease more (and perhaps more important) information than statistical significance. Some Simple Conversions to Move from Tests of Statistical Significance to Magnitude of the Relation or Effect Size.

Table 14.2: Conversions from Tests of Statistical Significance to Magnitude of the Relation

Conversions
$ES = \frac{2t}{\sqrt{df}}$
$r = \frac{T}{\sqrt{t^2 = df}}$
$r = \sqrt{\chi^2(1)/N}$
$ES = \frac{2r}{\sqrt{1 - r^2}}$

I mentioned there are many measures of magnitude of effect. Cohen's *d* and r^2, respectively, tend to be the most familiar and in use in psychology.

> Interpreting the effects is not always straightforward. A difficulty in interpreting measures is that different indices are not interchangeable.

For example, Cohen gave as a guideline for *d* of .2, .5, and .8 as small, medium, and large effect sizes. So the small, medium, and large effect size of Cohen's *d* would translate to an r^2 of .01, .06, and .14, respectively. And this illustration is with just two of the many measures of magnitude of effect and strength of relation indices. The lesson when reporting these is to add a statement that the effects are small, medium, or large in relation to conventional standards (e.g., Cohen). This might seem redundant with merely reporting the number but that may depend on which index you use and whether you elect one of the more esoteric (less frequently) index (e.g., η^2). If you are reading rather than reporting a study, it is easy to encounter a measure of effect size and not know how that translates to anything.

As a guide to one's own research, whenever possible include a measure of the magnitude of the effect or strength of the relation in addition to statistical tests.

> Report effect size even for effects that were not statistically significant because significance can depend so heavily on sample size.

Indeed, it is not odd for findings in a given study to not be statistically significant (e.g., $p = .10$) but with respectable (e.g., medium) effect size (e.g., $p = .10$, ES = .46; Simard & Nielsen, 2005). We have already mentioned that studies are underpowered and therefore not likely to detect differences with small or medium effect sizes. Thus, reporting of ES for all comparisons or tests provides a way to identify consistencies across studies.

In addition, meta-analysis discussed further below is often used to summarize and evaluate a body of evidence. ESs are used in these analyses as a common metric across studies that have used different measures. Hence, reporting ESs is important to facilitate integration of a given study with the larger bod of literature.

> Finally, as one reports ESs, decode the terms small, medium, and large if they are presented. That is, provide the quantitative metric used to make these designations, especially if the estimate is not the familiar Cohen's *d* or *r* or r^2.

For example, eta squared (η^2) is used relatively often, but few people know how the numbers translate to small, medium, and large ESs. These latter designations are still arbitrary but help orient oneself and the reader about the impact. Small, medium, and large effect sizes are mere guides; they are not special cutoffs and would be worrisome if they fostered yes or no thinking like statistical significance testing can do. Effect size (even—large) is not to be confused with the importance or practical value of an effect, even though it frequently is as I note below.

14.2.2: Confidence Intervals

ES (or some other measure of magnitude of effect) provides a point estimate, i.e., a specific value that estimates the population value. To supplement this estimate, confidence intervals (CIs) also should be used (Cumming, 2012; American Educational Research Association, 2006; American Psychological Association, 2010b; Thompson, 2008). *A CI provides a range of values and reflects the likelihood that the difference in the population (e.g., between groups) falls within a particular range*. The interval does not provide any certainty that the difference really is captured in the range. The range is based on estimates from the sample data and based on the same information that is used for statistical significance testing. Indeed, common values used for CIs are 95% or 99%, which parallel statistical criteria for alpha of .05 and .01. The formula for computing CIs was given in Table 14.1. As evident, *z* values used for significance testing (e.g., *z* score of 1.96 for $p = .05$) are used to form the upper and lower CIs.

CIs can be used to test for significance, but they usually are not used in that way. The test would be whether the CI range includes zero, which would mean that the null hypothesis of no difference (zero difference between means) falls within that range. But making a binary decision with CIs, if that is all that is done, is no different from a statistical test. The task is to take advantage of the range and the interval, i.e., how large or small it is as a measure of precision. The smaller the interval, the greater precision of the estimate.

CIs can be used with (computed for) both ES and the original metric of the dependent variables used in the study (e.g., scores on measures of depression or anxiety). For example, one could state that in mindfulness treatment A was better than mindlessness treatment B and yielded an ES (*d*) = .70, with $CI_{95\%}$: .35, 1.05. This means that if the experiment were repeated an infinite number of times, ES we obtained falls within the range of .35 to 1.05 by chance only 5% of the time. Alternatively, the same data may be presented as a mean difference (i.e., difference scores between groups 1 and 2 on some symptom scale such as the Beck Depression Inventory or the less well-known Kazdin You-are-More-Bizarre-than-you-Think Scale) as 15 points with (for example) a $CI_{95\%}$: 10, 20 points on that symptom measure. They are equivalent. The ES in standard deviation units is readily interpretable in terms of strength of effect (e.g., *a la* Cohen's

recommendations for small, medium, and large effects); the mean difference presented on the original metric of scores on a measure, with the CIs, communicates to those familiar with the measure of the range within which the differences fall on the measure.

CIs have been strongly advocated either as a supplement to or a replacement for statistical significance testing (see Coulson, Healey, Fidler, & Cumming, 2010). In either use, CIs are advocated here too because they move away from dichotomous thinking (significant, not significant) that null hypothesis statistical testing requires. The interval provides a glimpse of where the likely population parameters lie and whether—no effect might be reasonable too (i.e., if zero is in the range of that interval). Use of CIs is to be encouraged because of the additional information it provides. Even so, it is hardly a panacea.

> Among the issues, if power of the study is low, the CI is likely to be wide (large range) and that may include the possibility of little or no effect within that interval, i.e., zero would fall within the CI range.

So CIs do not solve other problems such as low power, but they do help in seeing beyond accept/reject the null hypothesis.

As a general rule, report CIs in studies in which you are conducting tests of statistical significance. This is the position repeatedly taken by professional organizations, journal editors, and scores of statisticians and methodologists. And as I mentioned, as statistical computation goes, CIs are so easy to compute and obtain. Perhaps because the relatively infrequent use of CIs over decades, researchers sometimes have trouble interpreting what they mean. For example, in one study, 330 researchers (from psychology, behavioral neuroscience, and medicine) were given different scenarios of results of experiments (Coulson et al., 2010). Among the key findings, many (60%) misinterpreted the findings of a study when both statistical significance tests and CIs were presented. A conflict was seen in the results when in fact both were consistent in showing an effect. Others also have reported that authors frequently do not know how to interpret CIs and make errors in describing what the intervals mean when they discuss their results (Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Thus, as straightforward as CIs are to compute and present that does not guarantee they will be interpreted by the investigator or consumer of the research. We have some more education to do among ourselves. It remains valuable to include CIs routinely in statistical tests. Part of a battle once was to use them instead of *p* levels and statistical significance tests. The more common view to guide is to include CIs because of the important information they provided about the range estimate of the population parameter rather than a specific point estimate.

14.2.3: Error Bars in Data Presentation

For many college students, error bars refer to places they went by mistake thinking that the prices for beverages would be lower for happy hour. Yet, the term has a slightly different meaning in statistics. When graphing data (e.g., bar or line graph), the means usually are presented for each group or condition. One can observe visually how different the means are on the dependent measure. Yet, means and mean differences are really not very interpretable without knowing some measure of variability. In fact, on any graph it is easy to distort visually how large a given difference appears. For example, mean difference of 10 points between two groups can be made to look very large or very small depending on the scaling (e.g., increments 1 or of 20, respectively, for each hash mark or 20) of the y (vertical) axis.

> *Error bars refer to a measure of variability usually plus and minus the mean that is actually graphed with the data.*

Different measures of variability are used (CI, standard error of the mean, one or two standard deviations) (e.g., Cumming, 2012; Cumming, Fidler, & Vaux, 2007). Arguably the best of the measure to use is the CI that plots the range of the interval based on 95% or 99% confidence, which reflects *p*s < .05 and < .01, respectively. This upper and lower limit of the bar is computed in the identical fashion as the CIs and are drawn on the graph.

Figure 14.2 provides a graph of hypothetical data with the error bar corresponding to .95% CI.

Figure 14.2: Hypothetical Data with the Error Bar Corresponding to .95% CI

Hypothetical data showing one group with the error bar computed for the 95th CI. Note the mean is the halfway point between the top and bottom of the interval as represented by the highpoint of the histogram.

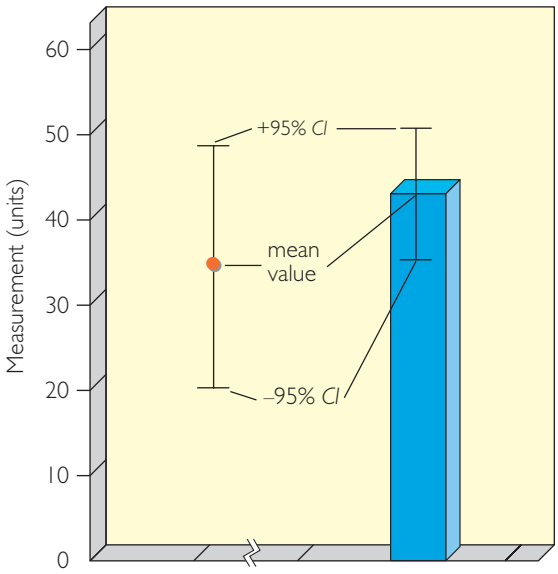
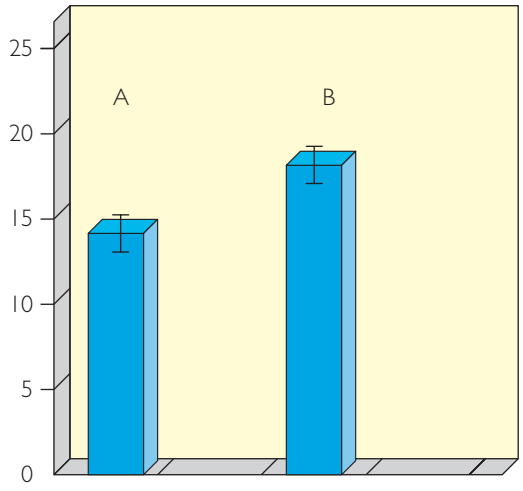


Figure 14.3 shows the means and error bars for two groups. By looking at the error bars, we can see if the ranges do not overlap.

Figure 14.3: Hypothetical Data for Two Groups

The bars represent means for each group. The vertical lines above and below the means represent the error bars drawn with the upper and lower limits of a CI_{95}. As evident, the intervals do not overlap suggesting that the group differences are statistically significant. Yet, more information is provided by seeing the extent of nonoverlap.



When the error bar ranges do not overlap, then the means are considered significantly different at least when CI_{95} or CI_{99} is used to compute the range. If there is overlap, this suggests the means are not likely to be statistically significant.

Yet, the error bars are not merely used to evaluate statistical significance. Indeed, I mentioned that different measures of variability are used that do not translate to the boundaries of statistical significance. The strength is in providing a range of values rather than just the means and to see the extent to which those ranges overlap between or among the groups and also the precision of the estimate (smaller range, greater precision in locating the population mean). Many of the comments made in relation to CIs are directly applicable here, but error bars are of course the visual equivalent when CIs are plotted.

As a general rule, when presenting data of means graphically, it is valuable to add error bars to the graph; note explicitly in the legend describing the figure exactly what the error bar means because different metrics are used (e.g., CI, standard deviation, standard error of the mean), and each requires a different interpretation; and consider CIs or standard error of the mean as the preferred metric (rather than one or two standard deviations above the mean) because of their connections with drawing statistical inferences in keeping with the prior comments about the strengths of CIs.

14.2.4: Statistical Significance, Magnitude of Effect, and Clinical or Practical Significance

Statistical significance and magnitude or strength of effect are now expected statistics in psychological research. In clinical psychology and other areas where the findings may be of benefit (e.g., psychotherapy, counseling, education, prevention), there also is interest in examining whether participants truly benefit from the intervention, program, or experience that was provided. This interest is reflected in the term "clinical significance" or "practical significance" of a finding. Clinical significance is the term frequently used in the context of intervention research and to refer to a change in treatment that may make a "real" difference or one that is important to clients who receive the intervention. Beyond a clinical context, it is useful to broaden the concept to *practical significance* to encompass all areas making a difference in real life important. Clinical is not the term in such areas as education, business and industry, and environmental studies because the focus is not on mental or physical health.

For example, to promote environmentally friendly behaviors, we might show that an intervention in one community increased the number of citizens by 5% who recycled their trash or who reduced their energy consumption in their homes, when compared to a control community that received no special intervention. The differences might be statistically significant, and the effect sizes may be large. Now we can look at the third criterion for evaluation that I am referring here as *practical significance* or whether the impact makes a difference in everyday life. Does that 5% gain make a difference in the community, have impact on the environment, or save money or resources in some other way? Statistical significance and magnitude of effect do not get at practical significance.

There is no single measure or index for practical significance. The metric may include everyday functioning or the quality of life of individuals, risk of some deleterious outcome (e.g., death), or the cost and cost-benefit of some intervention and a short-term or long-term index. The focus of the study (e.g., psychological functioning, disease, survival) determines what among many options would reflect impact beyond statistical significance or effect size. The three ways of evaluating the data, statistical significance, magnitude of effect, and practical significance are often confused. I have mentioned previously the significance fallacy, in which—significant statistically often is unwittingly interpreted by investigators as—significant, as that term is used in everyday life (i.e., to mean important or make a difference). This confusion is statistical significance and practical significance.

Much more common is the confusion of magnitude of effect (a purely statistical concept) and practical or clinical

significance (e.g., Kline, 2004; Prinz et al., 2013; Rutledge & Loh, 2004; Sun, Pan, & Wang, 2010). Here a large ES may be confused with a finding that has practical or clinical significance. Yet there is no necessary relation at all. One obvious reason is that the dependent variable that shows a large ES may be unrelated to everyday performance (e.g., a reaction time, variation in brain activation, specific cognitive processes on a decision-making task). Even if the measure is relevant to an applied problem, a large ES cannot be translated to practical significance.

For example, consider the results of a study for the treatment of obesity. We might recruit individuals 100 pounds (~45 kg) overweight and twice or more their ideal weight (criteria sometimes used to define morbid obesity) and randomly assign them to treatment or no-treatment conditions. All cases are followed up for 1 year after treatment ends. Assume that everyone in the intervention group loses 2 pounds (.91 kilograms) and everyone in the control group gains 2 pounds. Yet, at the end of the study all participants may still be very obese. Effect size for this result might be very large, but this does not convey whether the weight and health status have actually improved for anyone. Did treatment help anyone in palpable ways? We do not know from the ES, but a 2-pound loss is not very likely to qualify because the risk for serious health outcomes (e.g., diabetes, osteoarthritis, heart disease, or cancer) is unlikely to be affected by such a small change in weight. In other words, magnitude of effect as a statistical derivation has no necessary connection to clinical impact or practical significance. More generally, statistical significance, ES (or other magnitude of effect measures), and clinical or practical significance usually provide different information about the data, even though they are all quantitative methods of evaluating the results.

The previous example was one in which ES was large but practical significance or implications arguably were trivial. The other way shows the problem too, i.e., a small correlation can reflect an important practical effect. For example, individuals who experience clinical depression after a heart attack are at higher risk for dying in the following 6 months than those without depression (see Rutledge & Loh, 2004). More specifically, after a heart attack such individuals are more than four times more likely to die! Yet, using effect size equivalent yields a small correlation between depression ($r = .22$) and only a small amount of shared variance ($r^2 = .048$ or 4.8%). And the utility of aspirin in presenting a second heart attack has an important practical effect. Indeed, one clinical trial comparing aspirin and placebo was stopped early because the results were so clear in favor of aspirin, but the strength of the relationship was small ($r = .03$; see Rutledge & Loh, 2004). Several similar outcomes of this magnitude can still be quite important. In short, from the examples presented, ES or shared variance does not reveal importance or practical value necessarily at all.

Evaluation of the practical importance of the change usually is used as a supplement to statistical significance testing. In intervention research, once statistically significance (e.g., between intervention and control groups) is evident, further efforts are made to quantify whether the intervention has moved the client appreciably closer to adequate functioning, i.e., whether the change is important. Measures of clinical significance are of great importance. Beyond the context of treatment, measures that convey whether a finding makes a difference on a real-world measure are important. Psychology and psychological research is in the public domain and addresses many questions of interest to everyday life and policy. Consequently, beyond the usual research criteria (statistical significance, magnitude of effect), there is often great value in showing that impact of an intervention affects domains and measures that people care about (e.g., happiness, quality of life, optimism).

In research in applied areas or with foci on outcomes as in clinical psychology, counseling, education, prevention, medicine, and others, it is useful to go beyond the usual psychological measures designed to assess symptom change or improvement in a skill (e.g., reading), but to quantify in some way other indices to convey the importance of an effect. Two interventions could be different statistically and produce different ESs, but neither may have had much impact on everyday functioning of individuals or the groups that received the treatment (e.g., reflected in the number of days or years lost due to disability, as illustrated by disability-adjusted life year [DALY]). If the research one is conducting has applications for a problem (treatment, education, counseling), it is an extra strength to the study to include some measures that reflect impact that are close to if not directly reflect functioning in everyday contexts or metrics (e.g., costs) that have direct implications for use.

14.3: Critical Decisions in Presenting and Analyzing the Data

14.3 Report the presence of major decision points while doing data analysis

At this point we have discussed data analysis as a fairly straightforward process. Beginning at the proposal stage, one ought to have a good idea about how the data will be analyzed to test the main hypotheses. Once the data are collected, there will be other decisions based on such influences as attrition, variables differentiated groups and need to be controlled in the analyses, and unexpected analyses that emerged from intriguing or confusing findings. All of

this is routine science because all analyses cannot be planned in advance.

There are several decision points in conducting the data analyses. For many of these, there are no clear guidelines as to what the investigator ought to do. Yet, the decisions can have enormous impact on the results of a study and indeed whether a hypothesis is supported or not. It is important to be aware of these and to make decisions thoughtfully. Consider some key decision points.

14.4: Handling Missing Data

14.4 Identify some of the ways to manage issues that arise in statistical analysis when subjects drop out

Studies that are conducted over time require subjects to complete assessments on two or more occasions. For example, in longitudinal research assessments may be on several occasions spanning years. Similarly, but usually for a much smaller time period, intervention research (treatment, prevention) requires that subjects complete measures usually before and after the intervention, i.e., as characteristic of pretest–posttest design that was reviewed previously. In some lab studies too, the experimental manipulation and conditions will require participants to attend two or more sessions and complete assessments each time.

> In any study in which the subject is evaluated over time, there is the likelihood that some subjects will drop out of the study before all the data are collected.

Attrition or dropping out can threaten all types of validity, as we have discussed.

From the standpoint of the present discussion, loss of subjects raises special challenges for data evaluation (National Research Council [NRC], 2010; Twisk & de Vente, 2002). Designs begin by randomly assigning subjects to groups, and this is critically important to make implausible that selection biases could account for any group differences at the end of the study.

> If individuals drop out, the groups are no longer randomly composed. Stated somewhat differently, loss of subjects changes the study from a true-experiment to a quasi-experiment.

In a true-experiment, I noted that subjects are assigned randomly. That is the usual definition. But I did not mention the assumption that is inherent in true-experiments, namely, once assigned subjects remain in their groups. Letting some people drop out of one or more groups changes that. When anyone drops out, there is some unknown and difficult-to-document selection factor at work that leads people to say—I am not in this group (study) anymore. Even if there are only a few dropouts, there is a potential problem. I say potential because the number of dropouts in proportion to the number of people in the study may be relevant. In principle, one probably ought not to worry too much about losing 1 subject out of 5,000 but losing 5 out of 40, for example, now is actually rather than potentially worrisome.

Occasionally, investigators are falsely comforted by the fact that an equal or approximately equal number of subjects dropped out of each of the groups. Even if an equal number of individuals drop out from each group, this still alters the random assignment of subjects to groups and can create selection biases. That an equal number of subjects dropped out from each group does not mean that the same type of subjects or subjects with identical characteristics dropped out. Who drops out, i.e., their characteristics, may vary as a function of the condition to which they were assigned (e.g., one form of cognitive behavior therapy or medication rather than another or to a control rather than to a treatment group) and is not random. A key question here is how does one handle the missing data?

14.4.1: Completer Analysis

Three primary methods are used to manage missing data in studies with two or more assessment occasions (see Table 14.3 for a summary). The first is referred to as *completer analysis*. This has other names such as discarding incomplete cases and often in computer output listwise deletion. Authors usually state in their description of the subjects or beginning of the data analyses that some number of subjects did not provide complete data and hence were not included in the analyses. In laboratory with college students with one or two sessions, the subject who does not show up for the second (or later sessions) usually is just replaced with another subject. But in intervention studies and longitudinal studies, replacement in this way usually is not feasible.

> With completer analysis, the investigator merely analyzes the data for only those subjects who completed the study and who completed the measures on each occasion (e.g., pre, post).

Thus, in a two-group study comparing experimental conditions A and B, completer data analysis will only use subjects who have pre- and post-measures. The subjects without posttreatment (because they dropped out, died, failed to complete the measures correctly) will not be included.

Completer analysis seems to make sense. After all the prediction was that individuals who complete the experiment (go through the manipulation and finish the assessments at the end of the study) will differ from those who receive the other experimental or control condition. The

Table 14.3: Data Analyses of Treatment Trials When Some Subjects Drop out of the Study

Name of Analysis	Who Is Included	Comments
Completer Analysis	Only those subjects for whom there is complete data, i.e., they completed all of the assessments	The problem is that the random composition of the treatment groups is lost and threats to internal validity in particular. The groups can no longer be presumed to be equivalent except for the conditions (treatment) to which they were exposed
Intent-to-Treat Analysis	Include all subjects who begin the study whether or not they complete all the measures. For any missing data, use the previous data they have provided	This method preserves the random composition of both groups. The analysis can be a very conservative estimate of the effects of treatment. The reason is that subjects who dropped out during treatment will be included in the data analysis. The last data point they provided was the pretreatment assessment, and so these data will be entered at pre and post. Hopefully if there is a treatment effect, it will override the impact of cases who dropped out
Multiple Models to Estimate Missing Data	Multiple imputation models are applied to the same data using various assumptions about the nature of the missing data and using data from subjects to help with estimate the data points. The mean from the varied models serves as the estimate	This method preserves random assignment. The advantage is that different models that are used are not restricted to one set of assumptions (as is intent-to-treat) about the appropriate data estimate might be. There is no perfect way of estimating missing values but multiple models are likely to provide more defensible estimates

Each of the methods in the table is available from statistical packages commonly used by researchers (e.g., SPSS, SAS).

prediction is not that people who are assigned to a group but who do not receive the manipulation or only part of the manipulation will be different. Some people may drop out before receiving all of the conditions (e.g., if there are multiple sessions in a treatment study) or if they complete the conditions to which they were supposed to be exposed but do not finish the final assessments. Understandably, only those are analyzed who provide complete data. The reasonableness and seeming logic of completer analysis may be why this is the default method of handling missing data in many statistical software programs.

> Unfortunately, completer analysis maximizes bias in the data analyses, i.e., the groups are no longer randomly comprised and differences between groups may reflect who remains in the study or who responds to the specific treatment.

There is an internal validity problem (selection bias, selection *x* maturation—meaning the threat varied between the groups), an external validity problem (to whom do the results apply since those who completed treatment omit some set of subjects), a construct validity problem (was it the experimental manipulation alone or in combination with specially self-selected subjects who remained in their respective groups), and a data-evaluation validity problem (loss of power possibly based on who dropped out, changes in the means and variability in some way but not clear what ways). Other than these problems, everything is fine!

For a completer analysis, whether groups are or are not different from each other on the dependent measures, the results are difficult to interpret. Of course, so much of methodology is a matter of degree and hence common sense is needed here. If each group (*n*) consists of 100 subjects and there are two groups (N = 200) and 1 or 2 subjects drop out, it is unlikely that there will be bias. As the N gets smaller and the proportion of dropouts gets larger, selection bias is more likely to be a problem. In general, completer analysis is worth conducting as a complementary or supplementary analysis to one of the other strategies noted next. Yet, because it violates the random composition of groups, other strategies for handling missing data usually are preferable.

14.4.2: Intent-to-Treat Analysis

Completer analysis tosses out cases and that is one way to handle missing data. Another way is to engage in some method of *imputation*, *which refers to replacing missing values with some substituted value*. The most familiar way is referred to as *intent-to-treat analysis* and often is used as a way of handling missing data in clinical psychology studies. The procedure has other names such as *last-observation-carried-forward*, which nicely conveys how it works.

> Intent-to-treat analysis is designed to preserve randomization of the groups by keeping all of the cases in the study. This means that the data for any subject are analyzed according to the group to which he or she was assigned, whether or not the intended treatment was given, received, or completed.

Thus, even subjects who dropped out, whether at the end of treatment or in the first few minutes of the first session, or indeed, after they were assigned and never showed up again are to be included in the data analysis (Table 14.3).

Of course, if someone has dropped out of the study and does not complete the measures (e.g., at posttreatment or follow-up), how can one include them in the data analysis?

How do you think can one include them in the data analysis?

Typically, intent-to-treat analysis uses in place of the missing data the last (previous) data that subjects have provided. For example, if the study includes pretreatment, posttreatment, and follow-up assessment, we presume that all subjects have completed the pretreatment assessment. By the end (post) of treatment, some subjects have dropped out. An intent-to-treat analysis would include these subjects as if they had completed the treatment or served in the condition to which they were assigned. For those who dropped out during the study and who did not complete the postassessment, the pretreatment scores are used for the posttreatment scores as well, i.e., they are used in both places.

This sounds counterintuitive because these subjects did not complete treatment. Yet, by including all subjects the analysis will decrease any likelihood of selection factors to explain group differences. Methodological decisions and practices often are trade-offs and therefore it is valuable to know the arguments (or costs) on both sides.

> Intent-to-treat gives high priority to retaining the random composition of the groups, especially when compared to completer analyses.

Using the last (previous) data that subjects provided is the most commonly used method of intent-to-treat analyses. Another option is conducting an assessment of dropouts at the point that they drop out. That is, the posttreatment assessment battery is intended to be completed when treatment is finished and will be completed in this way for most subjects. For dropouts, sometimes one can obtain measures at the point they drop out. If subjects can be contacted and will complete the measures, the data can be used for their posttreatment assessment. Even though they have not completed treatment, they have a mid-assessment that will be used for their data, when analyses are completed at post. It is usually not feasible to collect data from dropouts, but this can be done (e.g., with monetary incentives, telephone rather than in person assessment interviews, and abbreviated assessment packets). With assessment that is more easily completed by the subject, the investigator hopes that the dropouts will complete the measures. Presumably, assessment completed at the point of dropping out is better than merely re-using the pretreatment data as a measure of pre- and posttreatment. Yet, point of dropping out assessment changes the time interval for all subjects—pre- to posttreatment may be 12 weeks for some subjects but 5 weeks for a few who dropped out but were enticed to complete the assessments then.

In the usual case, intent-to-treat uses the last data point without seeking any extra assessment. This method is often viewed as a conservative way of handling missing data.

It is assumed to be conservative because it implies that the treated group did not improve (pretreatment scores are used for the pre and then moved forward to be used at posttreatment as the same score). Yet this action is not necessarily conservative or bias free. Subjects may become worse from their prior assessment. The last data point in fact may be a poor estimate of the likely outcome and could misrepresent what postassessment would have been.

> There are subtle problems with moving forward the last number for the data analysis. Estimating missing values in this way changes variability.

Variability is reduced because the same number as a previous one is used to replace missing data when it is unlikely the real score would have been the identical number. Also, degrees of freedom are not quite right as the same number is used in two places and the observations are not independent entries. Nevertheless, intent-to-treat is relatively commonly used. The fact that it has become standard in many ways may be due to the ease of carrying out and explaining. Yet, among imputation methods it is not regarded as the method of choice because it makes assumptions about who drops out and why, whether the reason for dropping out makes moving the last data point forward is reasonable, and whether bias will be introduced (Carpenter & Kenward, 2008; NRC, 2010).

14.4.3: Multiple Imputation Models

Intent-to-treat is one way of imputing the missing values and uses one imputation model, namely, bring forward the last data point. A less used but better option is to use multiple estimates of the missing data. That is, within the data set, multiple models can be used to estimate what the missing data would be based on equations that draw on other data, including data from subjects without missing data.

> These multiple models might provide 5 or 10 estimates of what a missing data point will be. A mean estimate of the data point provides a more defensible estimate of the missing data in the general case.

Statistical software programs (e.g., SPSS, SAS) include multiple imputation models, although these models are infrequently taught and generally not familiar. The models essentially ask what would the missing data points be under different assumptions and assumptions that can be modeled or tested by the complete data of the other subjects as well as the available data of the subjects with missing information. The use of diverse models is beyond the scope of the present chapter because of multiple statistical considerations (e.g., parametric, nonparametric) and assumptions they entail. Yet as you consider how to handle missing data, recognize trade-offs, and consider different ways of imputing data. As I noted the amount of missing data is relevant too in relation to how much they influence the clarity of the conclusions.

14.4.4: General Comments

Missing data can be expected in studies where the number of sessions that participants attend (e.g., 2 vs. 5 sessions) increases, the duration of the study is extended (e.g., 2 weeks vs. 2 years), and the demands made on the participants (e.g., respond to a phone interview, come into the lab; complete self-report measures, give blood samples) increase. There is agreement about a few points related to missing data among statisticians and methodologists who have considered the matters (e.g., NRC, 2010).

When the study is designed, strategies ought to be planned to minimize attrition. The best plan is to develop specific ways to foster completion of the study. Among strategies often used are providing incentives for completion (money, chance at a lottery prize), minimizing burdens on the subjects (e.g., helping with transportation if needed), and keeping in close touch with subjects (e.g., holiday and birthday cards, newsletters about the project in a longitudinal study) in addition to any reminders about upcoming participation. In relation to missing data, the emphasis and priority ought to be its minimization from the very outset. There is no perfect solution to handling missing data. Each method has some trade-off and in terms of this text can influence some facet of experimental validity. Different methods make different assumptions statistically about the nature of the missing data, and these can only be tested in simulation studies (without—real data) and provide estimates of what the real data would have been like. There are fancy tests (called sensitivity analyses to help test assumptions about missing data), but these are infrequently used in psychology.

It is useful to track the reasons for dropping out. This can be very informative in relation to further research, leaving aside for a moment the matter of missing data. The reasons may generate new lines of work or modification of the experimental manipulation or intervention in future studies. Also, being able to report the stated reasons for dropping really improves the details of the study. We know much more than just that subjects were lost. In relation to missing data, some of the reasons may in fact be random or unsystematic (e.g., moving to a new city, death) whereas others are more clearly connected to treatment (e.g., did not care for the intervention, experience side effects) or to possible covariates of treatment (e.g., severity of the problem, age). Where there are large numbers of subjects and dropouts, how the missing data are estimated (imputed) can vary depending on who dropped out and why. That is, different imputation methods can model alternative ways of estimating the data based on random or nonrandom dropping out and lead to novel analysis that evaluates how robust intervention effects are in light how individuals dropped out.

As you can tell by now, it is possible to make a career out of the matter of how to handle missing data. As I noted no one method is correct and no one method is conservative. A truly conservative method would be to assign the worse outcome for a treatment group (i.e., insert the worst possible score in the opposite of the expected direction) and the best possible score for the control group (i.e., a number that shows the highest improvement possible). This conveys what conservative might look like.

Intent-to-treat seems to be the most common method used in clinical studies, even though, as I mentioned earlier, the method is controversial and not recommended as a blind or unquestioned strategy. As in many methodology issues, there is no single solution and it is important to understand what one is doing and the objections.

> Evaluation of the data with multiple methods (e.g., intent-to-treat and multiple imputation methods) is worth considering.

The question one is addressing by using different methods is, Do the results depend on assumptions made about the meaning (e.g., dropped out randomly or not) and estimation of missing data? Converging results from analyses of different imputation methods strengthens the case that the results do not depend on some hidden and not really justified assumption or actually that they do. Either way is informative and a better presentation of the study than just with one of the analyses.

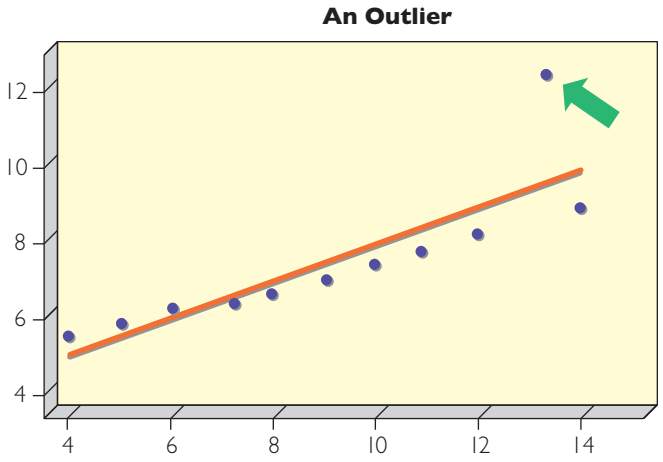
14.5: Outliers and the Prospect of Deleting Data

14.5 Investigate the rationale of deleting the outliers in a statistical experiment

An important issue that may emerge in data analyses is that some scores are rather extreme. *Outlier is the term used for an observation or score that departs greatly from the rest of the scores in the data.* The scores are not merely at the high or low ranges but are conspicuously separated numerically from the next nearest scores and have the function of distorting the overall distribution. Figure 14.4 gives one example where the graph plots the scores of individual subjects. Obviously, one's eyes are immediately drawn to the data point (subject's score) in the graph where one person scored quite differently from all other subjects. As a first step in data evaluation, it is quite useful to look at the distribution of all subjects graphically to better understand one's data (e.g., where subjects are bunched, whether there is skewing at one end or the other, and so on). Part of that is one can see if there are individual outlier possibilities or clusters of scores (subjects).

Figure 14.4: Plots the Scores of Individual Subjects

Hypothetical data plotted where each dot represents a subject's score. The extreme score (arrow) is an outlier, i.e., a score that departs considerably and conspicuously departs from those of the other subjects.



Consider an example in words rather than graphically. For example, 100 subjects in a study complete a measure that can range from 0 to 50. In this hypothetical sample, suppose that the mean is 20 with a standard deviation of 5. We assume a normal curve or distribution (available on a Web search or the appendices of many statistics textbooks), and from that we can tell what percentage of individuals is likely to fall within various standard deviation units. Thus, we can say that approximately 68% of subjects should have scores that fall within 15–25 (which is minus and plus 1 standard deviation of the mean) and that approximately 95% of all subjects should scores that fall within 10–30 (minus and plus 2 standard deviations from the mean). And now we see from examining the scores that there are two individuals pretty far out with scores of 38 and 39, which are over 3½ standard deviations above the mean. That is pretty far out because in a normal distribution only 1% of individuals are above or below three standard deviations.

> Clearly these two extreme scores can distort the data—they can raise the mean beyond the contribution of other scores because they are so high and they will add to the standard deviation (increased variability) because, as you know, the formula for computing the standard includes subtracting each score from the mean.

If we delete these subjects, it all becomes so neat and pretty. The distribution looks like a normal curve as opposed to the abnormal curve if we leave the data in and the results I wanted are statistically significant or more consistent once I toss the two odd subjects who were extreme.

This casual statement houses the concerns about outliers. One would like to have statistics that were not distorted by outliers and in describing a sample it is often useful to include means but also medians. When national statistics are provided to explain annual personal income, usually the median is used because that income is the one that is at the 50 percentile (divides the distribution of all income earners in half) and is not influenced by extreme scores (outliers).

> That is important because a number of huge outliers, people who make fortunes, or others who have absolutely no income influence the mean, but not equally. That mean (numerical average) is greatly influenced by outliers (e.g., the billionaires).

In statistical analyses in research, it is not quite that simple. But the rationale is the same, outliers may distort the data. Should they be deleted and, if so, by what criterion?

Some critical questions to raise. First, what is an outlier?

There is no standard answer that is mathematically or statistically justified or universally accepted, and many different definitions are used. That is important to know, so when you are making a decision there might be precedent but no clear justification. Actually, there are many definitions that have been used and many procedures to identify and deal with outliers. Among procedures reported, one occasionally sees the definition to include any score that is three or more standard deviations of the mean. Yet, a review of research identified over 20 different ways of defining and handling outliers (Arguinis, Gottfredson, & Joo, 2013).

Second, should outliers be deleted?

Statistical software programs offer options when computing data analyses on what scores to delete based on some outlier status. These options for deleting cases make the task easily accomplished with a click or two. Yet, why delete the outliers and on what basis? This requires much more careful consideration of what one is doing and why.

Finally, if one decides to delete outliers, by what procedure or methods and how will those different options alter the substantive findings and conclusions of a study? This leads to the worrisome possibility of publication bias as data analyses are completed with and without subjects who are identified as outliers and then reported with the analyses that confirm the hypotheses.

Several considerations can help answer these questions. Identifying outliers is one task and what to do about them is another matter.

Arguably, outliers ought not to be deleted because they merely distort the data. Outliers *are* part of the data. The reasons for their possible deletion are important.

If there is a strong procedural reason that can readily explain outlier performance that is important to note. As part of running the experiment, there may have been a mechanical breakdown (e.g., presentation of material, of

assessments) or a gross procedural error (e.g., running the subject in the wrong condition but identifying them in the data as in the other condition). These would be reasons to omit subjects, but these reasons do not have to do with the outlier status in their scores per se. There is less of quarrel about deleting subjects when gross errors in their participation were somehow involved. And that might be defensible independently of looking of the data they generated.

Also, it is important to check all data scoring and entry to be sure that all data were entered correctly. One number or two or a column or two on some database or data entry system can easily make a two-digit score (e.g., 10), a three-digit score (e.g., 100) or a small two-digit score (e.g., 10), a larger one (e.g., 80). In both of these circumstances, the investigator may be obsessing about what to do with an outlier whose data were not properly entered. That is, the subject may not be an outlier at all.

Consider that procedural and data entry issues are not at work and one has an outlier. What to do? Merely omitting them is controversial in part because outliers are—real data. That is just because their scores are not close to those of most people is difficult to justify. Also, using some criterion as departure in standard deviation terms or derivative methods makes the assumption that the population distribution is represented by the normal curve and the outlier is way out of bounds on that. Yet, there may not be a normal curve in the population on the dimension of interest and in fact all sorts of other distributions including two separate distributions that capture different kinds of people might be operating.

Indeed, a recent review of psychological research suggests that normal distributions are the exception rather than the rule (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013). The outlier then is not an outlier at all but part of the population data set that is real. Indeed, they can be exceptional people who make up an important part of life by virtue of being outliers, as discussed in lay textbooks that actually use that title (Gladwell, 2008). In addition, there are strong biological processes that actually foster diversity and sometimes large departures (outliers) from a distribution (e.g., Rebollo, Horard, Hubert, & Vieira, 2010). Thus, outliers might be anomalous in some ways, but that may not be justification for just tossing them.

Finally, and importantly, tossing data alters the randomization of subjects to groups. If a sample is reasonably large (e.g., 1,000) and a small number of subjects (e.g., 2) are outliers, their deletion is not likely to have much impact and raise selection bias as a threat to internal validity. Even so, before one tinkers with deleting subjects it is wise to look in the mirror and ask, Am I sure I want to tinker with randomization? (Each time I have asked this, I heard thunder and saw lightening.) If the sample is smaller (e.g., N = 50) and outliers are higher in number (e.g., n = 3), this is more iffy.

What should one do with outliers? Here are some recommendations:

Try to find a reason for outlier scores within the study (any procedural or data scoring issues) to make sure that those are ruled out;

Consider the method of handling outliers. Some methods (e.g., Winsorising) do not delete outliers but have a way to trim the data to retain extreme scores; Consider using more than one way of defining outliers if you are pondering their deletion. Do the different methods identify different individuals as outliers, or do they converge? Present the data analyses with and without the outliers so that you and readers of the write-up can judge the impact of any decision you made; Be completely transparent regarding how outliers were defined and what method was used to handle them; and Consider use of statistics (e.g., nonparametric) that can allow retention of outliers and provide tests of the hypotheses you wish.

As a general caution, be careful in throwing out subjects. The act is based on assumptions that often have no strong basis. So the subjects were extreme—they are real subjects. So the subjects were oddly different from everyone else. Come to one of my family Thanksgiving reunions (which we also refer to as our FOG [Family Outlier Gathering]). Extreme scores may be legitimate in athletics, these are often called world records; in personal income, we call it decadent wealth or abject poverty; in music, we call individual outliers—prodigies (e.g., Mozart); and in social skills, we call them jerks or emotionally intelligent—maybe emotionally brilliant. In other words, outliers are real and often all around us.

Two other issues about deleting subjects with extreme scores are important because they connect with other topics we have covered. The first issue pertains to extreme scores as a source of ideas for research. Studying individuals with extreme scores can be enormously valuable.

There may not be enough outliers to study them (e.g., 3 standard deviations above the mean), but studying extreme groups and in-depth studying of individuals who do depart can lead to significant breakthroughs in our understanding of broader issues and principles.

For example, effective treatments of cancer have come from finding only one subject who responded to a medication that otherwise was a failure, i.e., did not help people in general (Kaiser, 2013b). Yet careful analysis of that one outlier revealed several characteristics (e.g., genetic mutation of the tumor). The ineffective treatment proved to be effective when provided to others with this same characteristic.

More generally, in psychology the study of extremely shy children, in medicine the study of individuals with HIV whose disease does not progress to AIDs, and in genetics individuals or families with a rare disease—all of these and more have led to amazing breakthroughs. The expression of everyday life: The exception proves the rule perhaps is odd and merely a way to protect a cognitive bias about some

rule that is not actually very accurate or at least universal. Yet in science and methodology often it is true that the exception can inform the rule, i.e., help explain key processes or show that there are different rules for subtypes of individuals.

The second issue is to underscore that omitting subjects and those with extreme scores is a broad issue in science and not just psychological science. For example, nonhuman animal studies (e.g., mice, rats) on the impact of medication for disease and biological conditions occasionally introduce bias by deleting subjects that did not respond as expected. As one case in point, a drug designed to help the brain recover after stroke among mice began with 10 mice in the intervention condition, but only 7 made it into the data analyses (see Couzin-Frankel, 2013b). Further inquiry into those three that were omitted indicated that the mice had died. Not including subjects that died seems reasonable and perhaps obvious. Yet, the death was quite relevant to the study—they died from massive stroke, i.e., the brain was harmed by treatment. Omitting subjects changed the conclusions completely. In this situation and others like it, omitting subjects is not fraudulent or purposively misleading.

Consider another example. In this study, the purpose was to see whether the effects of sleep deprivation in adults could be detected on facial appearance (e.g., having more hanging eyelids, redder eyes, more swollen eyes, darker circles under the eyes, paler skin, more wrinkles/fine lines, and more droopy corners of the mouth) (Sundelin et al., 2013). Adults were assessed before and after sleep under two conditions—a normal night's sleep (8 hours) versus a night with only 5 hours of sleep followed by 31 hours of sleep deprivation. Of 23 subjects photographed and rated for fatigue, a subgroup of 10 subjects was selected and the photos of these individuals were evaluated. Yes, sleep deprivation showed that we look pretty much like one would expect—a variety of facial signs, which are all negative, are associated with deprivation. Yet, in this study three subjects were deleted; the reason, their photos did not vary from deprivation to nondeprivation conditions. Here again omitting subjects is arguable. There is no mischief at all on the part of investigators—there are no guidelines on how to handle many such situations.

Consider for a moment a curious way of highlighting the potential bias of deleting subjects. Just for the moment, consider an oversimplification, as follows: In a study, there might be two kinds of subjects—those who show the predicted pattern we as investigators expected, wanted, and prayed for and those who do not.

Which subjects do you think are more likely to be deleted?

The answer is the problem or can be a problem. This is not a mere external validity problem, i.e., we can only generalize to a subgroup of subjects who have some characteristics that led them to stay in the study or to be retained by us when we were omitting subjects. We have described before that attrition or loss of subjects (whether subjects drop out or investigators kick them out in the study or data analysis) can affect internal, external, construct, and data-evaluation validity.

As a more general comment, when you conduct research, examine the impact of your decision to retain or omit the data and be completely transparent so that the reader can judge what you have done and its implications. If one is reading research, look at the section that describes the subjects and see the sample size (N). Then read on (in the Method section or beginning of the Results) to find the—real N, i.e., the subjects who made it to the data analysis. Now find out what happened to those who did not make it to the data analysis and scrutinize why. The results might have been completely different depending on a variety of factors (e.g., how many subjects were omitted, why). Now a judgment needs to be made about the plausibility of impact on the conclusions.

14.6: Analyses Involving Multiple Comparisons

14.6 Examine statistical analyses that involve comparison of multiple subject groups

In an experiment, the investigator is likely to include multiple groups and to compare some or all of them with each other.

14.6.1: Controlling Alpha Levels

Here's an example of an investigator including multiple groups and comparing some or all of the groups. For purposes of this example, the study includes four groups—three experimental (A, B, C) and one control (D) groups. The investigator may conduct an overall test (analysis of variance) to see if there are group differences. If the differences are statistically significant, several individual comparisons may be made to identify which groups differ from each other. Alternatively, the investigator may forego the overall test. Several two-group (pair-wise) comparisons may be completed as each condition (A, B, C) is compared to each other and to the control (D) group. Alpha might be set at $p < .05$ to protect against the risk of a Type I error.

> This alpha refers to the risk for a given comparison, sometimes referred to as a *per comparison error rate*. However, there are multiple comparisons.

With multiple tests, the overall error rate or risk of a Type I error can be much higher, as a threat to data-evaluation validity.

The increase in Type I error rates from multiple tests is sometimes referred to as *probability pyramiding to note the accumulation of the actual probability of a Type I error increases with the number of tests*. How much higher the p level increases depends directly on the number of different comparisons. In fact, with a number of comparisons, each held at the per comparison rate of .05, the probability of concluding that some significant effect has been obtained can be very high. In our hypothetical example with four groups, the investigator may make all possible comparisons of the groups (six total pair-wise comparisons, A vs. B, A vs. C, and so on). Although the pair-wise error rate is .05, the risk of a Type I error for the experiment is higher because of the number of tests. This overall rate is referred to as the *experiment-wise error rate*. We must control for the probability of a Type I error for all of the comparisons or for the experiment-wise error rate. That is, the alpha selected must account for the number of pair-wise comparisons.

There are several multiple comparison tests that are available to address the problem of experiment-wise error rate and to control the increased Type I risk (see Bretz, Hothorn, & Westfall, 2011; Maxwell & Delaney, 2004). Many of the more familiar multiple-comparison tests are known by the name of the persons primarily responsible for their development (e.g., various tests by Tukey, Duncan, Newman and Keuls, Scheffé). A relatively simple alternative is referred to as the Bonferroni correction and consists of a way to adjust alpha in light of the number of comparisons that are made. Consider how the test operates. In a set of comparisons, the upper boundary of the probability of rejecting the null hypothesis is the number of comparisons (k) times alpha (α) (e.g., $p = .05$). Obviously, if there are 10 comparisons to be made then the overall error rate is $k\alpha$? or .50. As a protection against a Type I error, $p = .50$ would clearly be unacceptable. To control the overall error rate, alpha can be adjusted for the number of comparisons.

The Bonferroni adjustment is based on dividing alpha ($p = .05$) by the number of comparisons. In our four-group study, there are six possible pair-wise comparisons, as mentioned before. If we set alpha at .05, we know our risk is actually much higher given the number of comparisons. To make an adjustment, we divide alpha by the number of tests. In our example, we divide .05/6, which yields $p = .0083$. For each of the individual pair-wise comparisons we complete (group A vs. B, A vs. C, A vs. D, and so on), we use $p < .0083$ as the criterion for statistical significance. If we use this criterion, then our overall experiment-wise error rate is controlled at $p = .05$.

The Bonferroni adjustment controls the overall (experiment-wise) error rate, for example, at $p < .05$. The error rates for the individual comparisons (per comparison) need not be equal (e.g., all at $p < .0083$ in the prior example). Individual comparisons can vary in their per comparison alpha level, if the investigator wishes greater power for some tests rather than others, as long as the overall per comparison alpha levels do not exceed the experiment-wise error rate of .05 when summed for all comparisons.

> The adjustment of alpha, as noted here, arises when several pair-wise comparisons are made on a given measure.

A similar concern, i.e., elevated alpha, emerges when there are multiple outcome measures and multiple tests comparing the same groups for each measure. For example, if two groups of patients (anxious vs. nonanxious patients) are compared on several different measures, the chance of finding a significant difference, when there is none in the population, is higher than $p = .05$ for a given comparison. Here too, the Bonferroni adjustment can be used for the number of comparisons where k refers still to the number of comparisons or tests. As before, for each pair-wise test, the adjusted level is used to decide whether the effects are statistically significant.

14.6.2: Considerations

There is general agreement that multiple comparisons require some adjustment to control for Type I error. Failure to consider the multiplicity of the comparisons has direct implications for data-evaluation validity, in this case, often concluding that there are significant differences when, by the usual criteria for alpha, none exists. That is, when so many comparisons are made without controlling for the experiment-wise error rate, the likelihood of obtaining some significant effects by chance increases.

Beyond these general points, and at the point investigators need to make data-analytic decisions, agreement diminishes. For example, which multiple comparison tests are appropriate and whether a given test is too conservative or stringent are two areas where reasonable statisticians can disagree. Use of an adjustment such as the Bonferroni procedure is fairly common. Although the adjusted alpha is reasonable, the consequence can be sobering in a given study. In practice, the number of significant effects decreases when an adjusted level is used. Stated differently, as the alpha for individual pair-wise comparisons becomes more stringent, power decreases and the probability of a Type II error increases.

> Within the current practices of significance testing, control of Type I error, rather than Type II error and power, is given the highest priority.

Hence, investigators are encouraged (by tradition, research advisors, reviewers, editors) to keep alpha at .05 or .01 almost at all costs. The difficulty for this orientation in research is that we already know that power in most

psychological studies is likely to be weak for detecting small to medium effects. When adjustments are made to control overall alpha levels, power of a study decreases even further (because a more stringent alpha is used for individual comparisons). That is, apart from a relatively small sample size, the investigator is burdened by correcting for the number of statistical tests. Understandably, investigators are reluctant to adjust for the large number of tests they complete.

There are alternatives for the investigator who believes central findings are supported by the statistical comparisons but sees them disappear when alpha is adjusted to control the experiment-wise error rate:

1. The investigator can present the results for both adjusted and nonadjusted alpha levels. The results can note the tests that remain significant under both circumstances and those that are significant when left unadjusted. This is not a completely satisfactory solution, but addresses the ambivalence and tension both in the investigator and colleagues at large, namely, to identify what the effects are, to retain power at a reasonable level, but not to get carried away with an extraordinarily large number of tests, only a few of which are statistically significant. Also, transparency in what one has done and the differences obtained from various ways of looking the data are critical.
2. The investigator can select an experiment-wise alpha that is slightly more lenient than $p < .05$ such as $p < .10$ prior to making the adjustment. The Bonferroni adjustment will divide this alpha by the number of comparisons. The per comparison alpha is still below .05 depending on the number of comparisons. Adopting an experiment-wise rate of .10 is usually less of a concern to other researchers than adopting this rate for individual comparisons (per comparison rate).
3. The investigator may not be interested in all possible comparisons, but rather in only a preplanned subset that relates specifically to one or two primary hypotheses.

   Adjusting alpha for this smaller number of comparisons means that the per comparison rate (of alpha) is not as stringent. Indeed, for a few planned comparisons, not adjusting for the number of tests is usually viewed as satisfactory. Here, the difference is in conveying at the outset of the study what the hypotheses are and what specific tests will be used to evaluate them.

   Direct, planned, and a priori comparisons are usually favored. If any additional, supplementary, or exploratory analyses are conducted, these might be more conservatively tested (e.g., with adjusted p levels).
4. The number of tests might be reduced by combining dependent measures. Studies often include multiple measures to show differences between groups. Many of these different measures actually may be highly correlated and arguably can be measuring the same or overlapping constructs. As I mentioned, scores on measures that are highly correlated can be standardized (i.e., converted so that they are on the same scale) and combined into one measure. This actually can be done with more than two measures. In some of my own work, for example, we look multiple measures of therapeutic change among children referred for severe aggressive behavior. Although we care about individual measures, there are many and testing each one has all of the problems of multiple comparisons. We standardize main outcome measures, put them on the same metric, and analyze the single combined measure (Kazdin & Whitley, 2006). More generally, combining measures when they are correlated with each other can reduce the number of statistical tests. This alternative begins by looking at the correlations of all measures included in the study to identify if there may be high relations among some.

The alternatives do not exhaust the range of possibilities. One commonly used option is that investigators note there will be multiple tests and say they will adopt a more stringent criterion (e.g., $p < .001$ instead of .05) for individual tests before they call the effect statistically significant. This is better than ignoring the fact that there are multiple comparisons, but just selecting a more stringent p level is merely an unsystematic way to do the Bonferroni adjustment. The investigator (and reader) will not know the overall experiment-wise error rate. A better option is to be explicit about the number of tests and provide a clearer basis about how the adjustment was made in p levels to control for that number of tests.

We are discussing statistical significance testing, but another option is worth underscoring. One can deemphasize tests of significance altogether in the data analysis. Measures of the strength of the relation such as ES can be used and are not subject to the same concerns as statistical tests. This alternative was already elaborated earlier in the chapter. The central point is not to argue for any one solution but rather to underscore the importance of addressing the issue in the data analyses. Any data-analytic issue that can be anticipated also requires consideration at the design stage. Identifying the major comparisons of interest in the study, the statistical tests that will be used, and the number of tests may have implications for sample size and power. All such matters directly affect the conclusions to which the investigator is entitled and hence are critical to consider before the first subject is run.

14.7: Multivariate and Univariate Analyses

14.7 Compare multivariate and univariate analyses

In most clinical research, multiple measures are used to evaluate the impact of an intervention. For example, in studying self-control, emotion regulation, or other such constructs, several measures may be obtained to assess and evaluate the impact of an experimental manipulation on various indices of:

> Cognitive processes, rumination, affect, and perceptual and behavioral tasks.

Similarly, in therapy studies, several measures may assess the functioning in several domains (e.g., depression, self-esteem, adjustment at home and at work) and to rely on different assessment formats (e.g., interviews, questionnaires, direct observations). When there are multiple measures, the interrelations of the measures raise issues relevant to the data analyses.

Performance on several outcome measures may be conceptually related, because they reflect a construct (e.g., symptoms, well-being, adaptive functioning) or domain the investigator views as a unit, or empirically related, because the measures correlate highly with each other. For instance, if we have 10 dependent measures, we could analyze these separately with t or F tests. This would entail many tests of significance, and we would have an inflated error rate (beyond $p < .05$). We could address the problem of an inflated Type I error with the adjustment (e.g., Bonferroni) or by combining measures that are highly related, after standardizing the score for each measure, as noted previously. Usually measures that are related to each other are not combined into a single measure. The fact that the measures may be related raises another issue.

> Univariate tests, i.e., separate tests for each measure, do not take into account the possible redundancy of the measures and their relation to each other.

It is possible, for example, that two measures of trust or emotion reactivity show significant effects due to some experimental manipulation. The investigator may discuss how robust the effects are across two measures, when in fact, the high correlation between the measures argues for one construct rather than two. It is possible as well that neither of the measures shows a significant effect, but when viewed as a conceptual whole there is a significant effect. The measures individually may not provide as robust or indeed as reliable an effect as they do when combined.

> When there are multiple dependent measures, we can consider the data to be multivariate.

It may be desirable to conduct multivariate analyses (e.g., multivariate analyses of variance). Multivariate analyses include several measures in a single data analysis, whereas univariate analyses examine one measure at a time. We do not use multivariate analyses merely because we have several dependent measures. Rather, the primary basis is when the investigator is interested in understanding the relations among the dependent measures. The multivariate analyses consider these relations by providing a linear combination of the measures and evaluating if that combination provides evidence for significant differences. For example, the study may include three measures of anxiety. One multivariate analysis might be completed by combining these measures.

> If the overall multivariate analysis indicates a significant effect, this suggests that some combination of variables has shown the effect of the intervention or independent variable of interest.

After this finding with the overall effect of the multivariate analysis, one might then conduct univariate tests (individual F tests on each measure) to identify the specific differences on each of the dependent variables. As before, the alpha would need to be adjusted to avoid elevated Type I error. However, univariate tests may or may not show significant effects following an overall multivariate analysis. The multivariate analysis takes into account the relation of the measures to each other and evaluates the combination of measures. The univariate analyses ignore this facet of the structure of the data and may not lead to similar conclusions.

14.7.1: Considerations

It may be quite appropriate to analyze the multiple dependent measures with multivariate analysis or with several univariate tests. Multivariate analyses are particularly appropriate if the investigator views the measures as conceptually interrelated and is interested in various groupings of the measures separate from, or in addition to, the individual measures themselves. For example, there may be several measures of patient adjustment and family functioning. Within the study, the investigator may group all of the measures of patient adjustment and conduct a multivariate analysis to identify a combination for this overall conceptual domain and do the same for the measures of family functioning. Similarly, the investigator may be interested in broad classes of psychiatric disorders but also more specific disorders as well. For example, in a longitudinal study of twins, genetic underpinnings and early environmental correlates were evaluated for two broad classes of disorders: internalizing (e.g., major depression, generalized anxiety disorder, phobia) and externalizing (alcohol dependence, drug abuse/dependence, adult

antisocial behavior, and conduct disorder) disorders (e.g., Kendler, Prescott, Myers, & Neale, 2003).

> Different levels of analyses require multivariate tests in part because of conceptualizing domains as related. The individual tests of the components of these larger domains (individual diagnostic categories) are quite meaningful as well.

Multivariate analyses evaluate the composite variables based on their interrelations. This is a unique feature and is not addressed by performing several separate univariate tests. Separate univariate tests might be appropriate under a variety of conditions if the investigator does not view the measures as conceptually related, if the measures in fact are uncorrelated, or if the primary or exclusive interest is in the individual measures themselves, rather than how they combine or relate to each other. Investigators occasionally use the multivariate analysis as an overall test. Once significant, they proceed with several univariate tests. Usually, these latter tests are conducted with a per comparison alpha of .05, and hence the overall risk of Type I error is greatly increased. Findings of statistical significance here are a problem because the multivariate test was assumed to control for a Type I error at the level of alpha (p .05). The individual univariate tests, if conducted, still require consideration of the number of tests and the experiment-wise error rate.

14.8: General Comments

14.8 Report some key considerations on decision points in statistical analysis

I have addressed several facets of data analyses that often are behind the scenes. That is, they reflect decision points that are not directly discussed in relation to describing the data and drawing inferences from statistical tests. I have noted the key decision points (e.g., when to omit subjects, whether and how to impute missing scores). I have made some of these decision points explicit along with considerations that can guide the decisions.

Key points need to be made about the decision points:

1. Whatever the decision you make, be explicit and transparent about what you have done and why. Transparency of the process is critical.
2. Whenever possible select strategies you will use to handle the decision points before the study is run. As I noted not every analysis can be anticipated because intriguing or perplexing findings may prompt additional analyses.
3. Sometimes the decisions are reached based on looking at the data and seeing if the hypotheses are supported. That is, one might delete outliers and show the effects are significant but maybe based on the—cutoff for outliers. The investigator might select the option that leads to statistical significance on key comparisons.

14.9: Special Topics in Data Analysis

14.9 Evaluate the meaning of "exploring the data" in statistical analysis

Special topics in data analysis include understanding and exploring the data from one's study as well as conducting research based on previously collected data.

14.9.1: Understanding and Exploring the Data

Exploring the data from one's study arguably is a topic that could lead off the chapter. Presumably, a first step would be to understand one's data in depth before moving to hypothesis testing. Yet, I have had to place the section here because there are, so to speak, different uses of the term—exploring the data. One use is in the context of hypothesis testing and consists of beating the data to death with endless analyses to find statistical significance (may my dissertation rest in peace). That is, one—explores or more accurately—searches with the goal of finding which combination of subjects (outliers in or out), measures (which measure will I include or not), and analyses (not all statistical tests of the same data and hypotheses yield the same results). This is post-hoc exploration fishes for statistical significance. I mention this to get this out of the discussion. Exploring the data is not about fishing, engaging in practices like that, or hypothesis testing even.

> Exploratory data analysis is an effort to understand the data and hidden structures, information, and other facets that might well generate hypotheses.

Here we are doing analyses to optimize what we can learn from the data. It is useful to begin by stating—we have completed the data analyses for the study and have tested the hypotheses and maybe have even written up a draft of the results. I state all of that because this exploratory analysis is not identified to find significance no matter what. It is to look for relations among variables and hidden structures in the data that may be interesting, odd, surprising, or disappointing. The goal is to understand as much as possible from the data set and indeed see if nuggets you find can guide the next study. That is better said by noting that exploring one's data is offered in the context of generating hypotheses (for future work) rather than milking variants of the data to test a hypothesis.[1]

The rationale for this latter type of exploration of the data is easy. We have collected this data set and are testing a few hypotheses. With that done, we would like to find any interesting relations somehow embedded in the data. As I have noted before, it is unlikely I will do a study in the same way ever again. I want to utilize the data set of search for something that might be of interest. If you have little theories here and there all the better, but here I admit to crass empiricism. So if some subjects responded as predicted, I would want to look at those who did not respond. I might compare these two groups and for myself use a lenient alpha level ($p < .20$). Mind you, I am not claiming these results are important and my peeks at the data maximize experiment-wise error rate, but I am just looking for possible leads to make me think about the findings in novel ways. I am looking for the expected and unexpected. The explorations may suggest a moderator to explore in future research or challenge some view I had about possible mechanisms. Or I might find a strong set of correlations and say to myself, These make no sense. When my views (make no sense) conflict with the data (the correlations are there, strong, and in my face), I go with the data. The challenge is to consider why the relations might make sense and probably there is an intriguing study that follows.

My comments are about going beyond tests of hypotheses and to look or other processes, outcomes, and connections that might identify relations and generate research. My comments are informal, but it is important to note in passing that exploratory data analysis (EDA) is a formal framework for exploring data that involves statistics, graphing, and modeling and has a long history (e.g., Tukey, 1977). There are several tools that help reveal patterns in the data that would not otherwise be evident, including multiple and novel ways of graphing the same data or subsets of the data (e.g., scatterplots of various types, mosaic charts, network graphs) and examining and modeling of so called—error (residuals).

> It is important to be aware that there are formal ways to explore one's data. Even without training in some of the sophisticated methods, it is useful to explore the data well beyond the usually restricted tests of statistical significance designed to evaluate specific hypotheses.
>
> The guiding questions might be in this study, what happened to whom, how, with what exceptions? Correlations, partial correlations, tests of supposed subgroups, graphing the data in different ways or for different groups, all might generate novel ideas for your next study.

14.9.2: Research Based on Previously Collected Data

Much of research consists of running subjects—and that has consisted of direct contact with the human or nonhuman animals who are participating in the study. Even that is very much evolving as an increasing proportion of studies are conducted online and subjects participate from the privacy of their own computers and homes. Leaving that aside, much research is conducted without running subjects at all and this too promises to increase in the coming years.

> *Secondary analyses as that term is used here refers to conducting empirical studies based on data already collected and available.*

That is, one does not run subjects in the sense of collecting new data, but rather draws on available data sets. Consider two broad scenarios, the first of which is very familiar.

> **Meta-Analysis**. Decades ago, one reviewed the literature by reading all research possible and integrating the information in a narrative format. These were reviews of the literature and consist of combining and integrating studies and extracting from that conclusion about the current status of some body of literature.

These narrative or qualitative reviews are still frequently published in review journals (e.g., *Psychological Bulletin; Clinical Psychology: Science and Practice*). Yet there was a major sea change now some almost 40 years ago in clinical psychology. Major quantitative reviews of psychotherapy were published and introduced a new way of reviewing the literature, namely, meta-analysis (Smith & Glass, 1977; Smith, Glass, & Miller, 1980). The effects of psychotherapy had been reviewed many times, but the meta-analysis provided a way to quantify treatment effects across hundreds of studies and to examine and explore variables in novel ways.

Meta-analysis consists of a methodology that is used to combine the results from different studies. Typically, the analysis is completed by converting measures from individual studies to effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009; Valentine, 2012). Thus, no matter what the different measures are in the original studies, they are converted to a standard metric, usually effect size (ES). With a common metric, now the studies can be combined. With many different ESs and many different studies and meta-analysis provides a better estimate of the population parameters (true effects) than any one or two studies might provide. Also, one can form CIs as well to provide the range in which the population parameter exists is more precise. In short, an initial feature of meta-analysis is a way to combine studies quantitatively and to get estimates of the effects across many different studies. This feature eliminates much of the inherently subjective feature of a narrative review in which a reviewer may feel she is comparing apples, oranges, and socks, because the studies are so very different. From such diversity, it is difficult to rise above the variation to draw general conclusions and then conclusions that have any precision. Meta-analysis helps along these lines tremendously.

Meta-analysis is fairly familiar now that thousands are available in the psychological literature. Yet it is useful to briefly note the standard steps for conducting a meta-analysis (Card & Casper, 2013). The five steps are:

1. To identify the articles to be included. This requires defining what will be the topic and search terms. Cognitive behavior therapy may be the focus, but there may be many such terms that can be used to be sure as many studies are identified as possible.
2. To devise inclusion and exclusion criteria. Of all the articles that might be identified, are there articles one wants to exclude (e.g., dissertations, studies with children; studies that focus on non-clinical samples, all studies or only RCTs)?
3. To decide how the search will be to capture the pertinent studies. Typically, the search words are applied to multiple electronic databases (e.g., PsychINFO, Medline, ERIC for psychology, medicine, and education, respectively). Yet, these databases miss many studies, and it is useful to supplement electronic search by going to pertinent journals that often publish on the topic and with studies likely to meet the criteria, looking at recent articles that are identified and searching their reference sections to identify other studies.
4. To devise the coding systems to categorize and evaluate the body of research. These codes reflect the variables of interest to the investigator. These variables may be useful for describing the literature (e.g., how many studies included this or that population, evaluated the impact of the intervention on daily functioning, assessed follow-up). Yet, the more useful focus is on looking at the possibility of moderators. Are the effects of the intervention or manipulation different as a functioning of the types of measures that were used? Does the type of client make a difference in outcome? Does quality of the methodology or do specific practices (e.g., random assignment) make a difference? And so on. These codes are based on the interests of the investigator conducting the analysis. The codes become the independent variables and are evaluated on the dependent variable (ES). Each study will be coded on all of the dimensions of interest. Consequently, the codes need to be checked to make sure that they are reliably used by different raters.
5. The ESs are calculated from the studies. And this can be done in different ways, and decisions here can make a difference. The size of the sample of individual studies can introduce a bias in the estimates of ES, so the computations use an index that takes that into account (see Cooper, Hedges, & Valentine, 2009). Finally, the analysis now describes (summarizes the literature) and also tests whether the moderators (codes) influence the outcome (ESs).

> As a critical feature, meta-analysis goes well beyond merely summarizing a literature in a quantitative way. Questions can be asked of the literature in a meta-analysis that an individual study did not or could not easily provide.

These are the moderators mentioned previously. None of the moderators studied by the meta-analyst may have been included in any individual study. But by looking across multiple studies, one can find the variation needed (e.g., young vs. old clients, etc.). This feature is the reason to raise the topic of meta-analysis in relation to secondary analyses. Using meta-analysis, one can conduct original research by asking questions of the literature that were not addressed and often could not be addressed by individual studies.

For example, a recent meta-analysis evaluated the role of the therapeutic alliance across many different treatments, clinical problems, and types of measurement (Flückiger, Del Re, Wampold, Symonds, & Horvath, 2012). Alliance (the bond between the therapist and the client) has been studied in literally thousands of studies. The meta-analyses allowed evaluation along multiple dimensions that could not be studied in any individual study. Studies were coded for a variety of variables that might serve as moderators of the effect of alliance (e.g., whether the study was a randomized controlled trial, whether structured manuals were used, whether specific or more general treatment outcome measures were used). Alliance influenced treatment outcome, a finding suggested in prior research, but the meta-analysis showed the effects were ubiquitous across many different types of treatments and were not moderated by the variables included in this study. The search for the effects of specific moderators conveys the original research opportunities that meta-analysis provides. The studies are already done; the meta-analyst can adopt creative hypotheses about how effect sizes may vary or differ as a function of other variables. Also, one can study the role, utility, and impact of characteristics of the research design meta-analysis. For example, in this meta-analysis whether a study was a randomized trial or did not make a difference (moderate) on the impact of alliance.

Needless to say, meta-analysis in clinical psychology has many uses well beyond evaluating interventions. For example, recent meta-analyses focused on:

- Neuroimaging studies to evaluate and test alternative models of what processes in the brain and what brain centers are likely to be involved in emotion regulation (Buhle et al., 2013);
- Performance on emotional tasks of individuals with clinical depression (Epp, Dobson, Dozois, & Frewen, 2012);

- Variations among cultures in relation to locus of control and symptoms of depression and anxiety (Cheng, Cheung, Chio, & Chan, 2013);
- The changing nature of social support and life events over the course of adult development (Wrzus, Hänel, Wagner, & Neyer, 2013).

In short, there is no restriction at all of meta-analysis on intervention, despite the excellent use in that context. The method is useful when one wants to summarize a literature quantitatively and to test novel predictions that usually could not be or were not evaluated in the constituent studies. With meta-analysis, high-quality and novel research including tests of theory, mechanisms, and moderator can be completed without ever running a subject.

> **Secondary Data Analyses**. *Secondary data analysis refers to the analysis of data that were collected by someone else.* Meta-analysis might well be subsumed under this, but secondary data analysis has its own use, meaning, and methods.

The analysis recognizes that there are many large databases and data sets that are rich and readily available. From these questions can be asked and addressed that are well beyond those for which the data may have been collected.

The data sets can come from many different sources. Primary among these sources are government agencies (e.g., federal, state) but other sources as well (e.g., university records, supplementary material from journal articles that required data and code materials). There are scores of national surveys and data sets available as conveyed by just a few examples:

- The Centers for Disease Control and Prevention (CDC) has multiple databases available on child and adolescent health, health risk behavior, maternal and child health, birth defects and developmental disabilities, and many more (www.cdc.gov/surveillancepractice/data.html).
- The National Database for Autism Research, through the National Institutes of Health, provides a repository of data available for secondary analyses on autism spectrum disorder (http://ndar.nih.gov/).
- The Human Genetics Initiative by the National Institute of Mental Health provides access to family data for studying the genetics of schizophrenia, bipolar I disorder, depression, Alzheimer's disease, autism, obsessive-compulsive disorder, and other mental disorders (www.nimhgenetics.org/access_data_biomaterial.php).
- The National Comorbidity Study focuses on the epidemiology and course of psychiatric disorders and makes available multiple large data sets.
- These are only a few samples merely to convey the overall point that rich sources are available. Searching the Web for databases for a particular disorder or type of behavioral problem or searching government agencies beyond those noted above can yield scores of databases.

The use of databases for secondary data analysis is not merely downloading a data file. For example, one project focuses from the CDC on School Health Policies and Practices. This is a project that addresses diverse domains (e.g., physical education and activity, mental health, nutrition, social services, health education, and more). The data for this project are available for secondary analyses. However, the data for 2012, for example, include files for 13 questionnaires, 15 data files, and 15 codebooks (and many others) along with additional instructions on how to use the database. The data are made available in multiple file formats, and programs are offered to convert the data from one format to the other.

Where to begin for secondary data analyses? The steps for conducting secondary data analyses have been nicely articulated (Donnellan & Lucas, 2013) and presented in Table 14.4. These steps underscore the importance of

Table 14.4: Steps for Conducting Secondary Data Analyses

Steps	Description
1. Finding Existing Data Sets	There are multiple sources with government agencies (e.g., National Institutes of Health, Centers for Disease Control and Prevention); also many projects in clinical psychology, psychiatry, and epidemiology are well known because of the large publications that already have been conducted by the primary investigators. The National Comorbidity Study is one such example and includes multiple large-scale data sets for secondary analyses http://www.hcp.med.harvard.edu/ncs/
2. Read the Codebooks	Mastery of the original procedures, measures, and data description and prior analyses are essential. The documentation is critical because unlike primary research where an investigator has designed the procedures, in secondary research one has to catch up with what was done and why and how that may affect the new study
3. Acquire the Data Sets and Construct a Working Data File	Obtaining and using the data may have restrictions or require special requests. Institutional Review Board (IRB) issues might be involved. However, when datasets are made available, the goal is not to introduce obstacles. Once the data are obtained, it is useful to develop small data files to focus on the study of interest for the secondary data analysis. The large file is a useful base to retain but may be too cumbersome to work with.
4. Conduct Analyses	This could be straightforward but also could have surprises if some variables are coded (reverse coding) in ways that were unexpected. A useful point of departure is to conduct analyses that repeat those that were conducted by the primary investigators just to make sure that the data (e.g., measures of central tendency and variability) are—behaving the way they were in the original studies.

in-depth understanding of the procedures and methods. Yet, the investment may be especially worthwhile because for an investigator, the database may serve as the basis of more than one investigation.

There are many advantages to conducting studies on such data sets. Among them are the likelihood that:

- The project is a large scale and includes data that are extensive at a given point in time and over time. This means that many constructs and measures were included.
- Many of the data sets are cohort studies in which one can establish the time line between early variables and later outcomes.
- The sample in such data sets is often special in one of two ways:
  1. The sample may be representative of a population of interest, and careful sampling methods were used to represent the sample or to oversample as needed to ensure all groups were represented. Psychology rarely uses random selection in selecting subjects for research.
  2. Many projects include special samples (twins, all individuals born at a given point in time, children with autism) that are very difficult to obtain or to accumulate in large numbers without a well-funded project conducted by many investigators from multiple sites.
- The research involved an extensive (and expensive) team in the design of the study, recruitment and maintenance of the sample, and collection and analysis of the data.
- The costs of each of the above features in grant funds and personnel are enormous and could not be easily replicated.

There are challenges as well. One must come up with an important question that can be answered by the data and of course that has not already been answered. Also, the original investigators may have made many decisions in selecting measures and how they are scored, used shortened forms, computed new variables, made critical decisions along the lines we have discussed (e.g., handling missing data, outliers), and all of that is pertinent to evaluating the data in any new way. The time—saved in actually running subjects is compensated for by making sure that one understands details of the data collection, summarization, and so on. Just because a large-scale study has been completed does not mean the project is free from questionable methodological practices. For example, there might be some scales devised by the investigators that are a few items summed to reflect a critical construct. All that said, secondary data analysis is mentioned here as a viable research strategy and one with special data evaluation issues and challenges.

Many disciplines (e.g., economics, political science, sociology) use secondary data analyses. Psychology has shown some reluctance in part because there is a strong experimental tradition of developing hypotheses and running subjects to test them, i.e., primary rather than secondary research (Donnellan & Lucas, 2013). In some graduate programs in psychology, for example, there are formal or informal restrictions on analyzing pre-existing data (e.g., for thesis or dissertation requirements) and individuals are strongly encouraged or required to run subjects. Of course, one might argue for a higher order criterion, namely, students and all of us should be trained to do the best science possible (e.g., use of theory, strong tests of hypotheses, important questions) and different paths can be used to get there (e.g., running or not running subjects; quantitative or qualitative research, etc.). And it is useful to keep in mind that some great scientists, perhaps as operationalized by receipt of a Nobel Prize, did not seem to run subjects (e.g., Einstein) but others did (e.g., Pavlov).

> Secondary data analyses are likely to increase in the coming years. There is increased interest in making data available to bring together larger databases as repositories on which many scientists can collaborate and draw. The rationale is that scientific progress can be more rapid if data are shared, combined.

Also, there is a new form or qualitative leap in secondary data analysis called—big data that represents sharing of material and raises novel issues. Yet, it is also a secondary data analysis in keeping with the points noted here.

Second, the ability to store information on a large scale and to make that information available has expanded the ease of providing data and all of the supplementary materials to make sense of how they were collected and analyzed. Hardware, software, and storage (e.g., on the cloud) allow for much more extensive data and all the coding materials needed to evaluate that data. The information can be accessed internationally in ways not previously available.

Third, data sets and banks of information bring together that is difficult to accumulate. For example, there is a keen interest in the psychiatric and physical damage that football players suffer from repeated head contact, injury, and concussion. Small-scale studies (a handful of retired professional football players) have revealed untoward damage in critical brain regions and processes (e.g., Small et al., 2013). More comprehensive data sets are available, and there is an autopsy database in process to better understand the scope of impact of head contact among professional football players and from that means of early identification and prevention.

Already brain autopsies are now readily available and are of keen interest in studying psychiatric disorders (e.g., Deep-Soboslay et al., 2011). These—psychiatric brain banks permit the accumulation of the data and set common

methods and standards to facilitate use for research. The information is such that would not be available by any individual investigator or team.

Overall, secondary data analyses provide rich opportunities. There is a way in which primary research is easier. In this, the usual case, the investigator develops the hypotheses and designs the study (experimental arrangement, appropriate sample, measures) to test the hypotheses. Secondary data analyses begin with the fact that key decisions have already been made and settled. The data are in and now one must as it were work backwards from that to creative hypotheses that can be tested. Occasionally, secondary analyses are advocated for individuals beginning their careers because one can utilize data that otherwise might take years to collect and with resources the early career investigator may not have. However, for present purposes, who conducts the analyses is immaterial. The issue is that rich data sets are available, and one can conduct high-quality research based on their utilization.

Summary and Conclusions: Presenting and Analyzing the Data

The chapter began by discussing basics before any data analyses are completed. This included careful checking of the data at all points where errors might be introduced. From that point, one begins the preliminary analyses to describe the sample and the groups. Among the goals is to evaluate whether any unanticipated differences emerged between the groups that might need to be considered in the data analyses. If new (home-made) measures are included in the study, preliminary evaluation of these measures is warranted too.

The main feature of data analysis is the use of statistics to draw inferences about the experimental manipulation. There has been ongoing dissatisfaction since statistical significance testing emerged about the utility of this approach for research. Among the many concerns is the fact that null hypothesis and statistical significance testing give us arbitrary cutoff points to make binary decisions (accept or reject the null hypothesis), and most importantly do not provide the critical information we would like (e.g., direct tests of *our* hypotheses and information about the strengths of our interventions). This chapter presented information to supplement statistical significant testing. Among those is the inclusion of a measure of the strength or magnitude of the relation whenever significant tests are presented. Effect size (e.g., Cohen's *d*) and Pearson product-moment correlation (*r*) are two commonly used measures, but there are many others that are readily available. Also, any point estimate such as effective can be greatly supplemented by providing CIs or a range of values about that effect. When graphing data, error bars provide another way to present the range of values that facilitate interpretation of an effect.

In conducting statistical analyses, there are several decision points. These decision points have no firm guidelines but how the decisions are made can greatly influence the findings that are reported. Completer analyses, intent-to-treat, and models of imputation were discussed as ways of handling missing data in research where there are repeated measures and cases that drop out before all measures are completed. Outliers in the data and deleting data were also discussed. Also, multiple comparison tests and the need to control error rates (Type I error) were presented as well. Finally, the uses of multivariate and univariate tests and the relation of these to error rates were highlighted.

Special topics in data analysis were covered and began with understanding and exploring one's data way beyond the interests in testing hypotheses that guide the study. Exploratory data analyses are designed to look at and explore a variety of relations and special findings that might generate hypotheses for future research. Exploration has the goal to understand and go beyond the direct tests of the hypotheses and not, of course, to search for significant findings that can be reported as—exactly what was expected.

Meta-analysis was also discussed. This now is a commonly used way of combining studies to summarize a literature but also to ask novel questions that usually go beyond what any single study has focused on. Thus one can look at moderators across studies or methodological features of the studies. Characteristics of interest to the investigator are coded for each study and serve as independent variables. Effect size, which is the common metric that is derived from the studies in the analysis, becomes the dependent variable. The common practice of meta-analysis as a way of evaluating a body of research on a given topic is yet another reason to argue for the routine inclusion of ES in any empirical study. The use will allow others to better integrate the study into a larger body of literature.

Finally, secondary data analyses were discussed, i.e., completion of studies that are based on data that other people have collected. There are rich databases that provide special opportunities for further evaluation beyond the original goals of the study. Secondary data analyses have their own

advantages and challenges. They are likely to increase in the coming years as funding agencies are making such data sets increasingly available and because much of the data collected are intensive efforts (careful and extensive assessments) with special populations accumulated in large numbers (e.g., children with autism, patients with schizophrenia, football players who have donated their brains for further study). Clearly data such as these provide special opportunities.

Among the messages of the chapter is that statistical analysis is not merely a matter of applying some technique to help draw inferences. There are many options, assumptions, and points of decision making, and these materially affect the conclusions. It is important to be thoughtful about these and then to convey the benefits of that decision-making process as one describes one's results. Failure to replicate a study can easily happen by failing to understand the data-analyses decisions and what was done at critical junctures (e.g., missing data, deleted data).

Critical Thinking Questions

1. What does each of these reflect in a research report: statistical significance, magnitude of effect, practical or clinical significance?
2. Outliers, individuals with extreme scores in a study, sometimes are tossed out by an investigator, sometimes they are retained. Give an argument for each side of this issue. What is your view of this?
3. It has been difficult to get investigators to use confidence intervals (CIs) in their reports, even though this is so easy to compute. Part of the problem seems to be that many investigators when surveyed are not sure what CIs really are. Give a dazzling definition of what a CI is.

Chapter 14 Quiz: Presenting and Analyzing the Data

Chapter 15

Cautions, Negative Effects, and Replication

Learning Objectives

15.1 Express the importance of using the right language to communicate statistical results

15.2 Explain why no-difference findings occur in scientific research using the five reasons

15.3 Analyze the utility of the negative result in statistics

15.4 Define the concept and role of replication in statistics

15.5 Explain why replication is important in scientific research using the five key reasons

In many ways, methodology is all about interpretation of findings in a study. As scientists we engage in special methodological practices, so the results can be interpreted in one way rather than another. Thus, we want to interpret the findings by explaining how a particular variable of interest to us, rather than some other influence, artifact, or bias (e.g., pre-existing group differences, "chance") is the basis for the results. Also, when the data are collected and analyzed, we want to explain the results in ways that are consistent with what we actually found. Often an investigator makes a little leap moving from the data analysis to the interpretation of what was found. The study then is revealed to be poorly designed. That is, in reading a report of the study, we say, "If this is what the investigator wished to conclude then this was not quite the right way to design the study." Thus, data interpretation issues are squarely within the realm of methodology. Indeed, it is helpful before designing a study to know exactly what you would like to conclude if your theory or hypothesis is supported. The design is built around making it so that you can reach that conclusion.

This chapter discusses interpretation of the findings. The focus is on the findings of an investigation and common issues and pitfalls that emerge in moving from describing and analyzing the results (and Results section) to interpreting those results (Discussion section). Also, the chapter focuses on so-called negative results, i.e., the absence of differences. Not finding statistically significant differences in a study is often viewed as non-informative and "negative." There are many exceptions to this and how and when so-called non-effects are important are elaborated. The final topic of the chapter has to do with interpretation that extends beyond one study and focuses on the critical issue of replication of research findings. Replication or reproducibility of research findings is so central to all of science and is one of the stronger protections we have against all sorts of potential and real problems (the finding of chance effects, biases in the study we cannot detect, and even fraud). Yet, there are strong forces that interfere with replications as a corrective factor, as we discuss.

15.1: Interpreting the Results of a Study

15.1 Express the importance of using the right language to communicate statistical results

Data interpretation has to do with how to talk about and discuss the results. I focus on communication of the results of a study and writing up the findings. In the write-up of most studies, the main sections are Introduction, Methods, Results, and Discussion with all sorts of subsection headings as needed. The Results present the quantitative evaluation of the findings with basic and fancy statistical tests usually (e.g., *p* levels, ESs, *r*s, *t*, or *F* test numbers). The Discussion section now steps away from the quantitative analysis and *describes* and *interprets* those analyses in narrative form. That narrative form

only means that we now place into words what the findings are without using any of the numbers. Description and interpretation of the findings look like quite different concepts and tasks and at the margins they are. Description (just say what happened) and interpretation (why did that happen) can be readily distinguished. Yet sometimes description is a little leap from what was found and includes some inferences or interpretation. These leaps are about methodology, namely, going beyond what the design, assessments, results, and other features of the study would allow. My comments here are about the discussion of the quantitative results.

> Data interpretation can be tricky because the meaning of the quantitative results of a study can be easily misinterpreted and overinterpreted.

Moreover, mis- and overinterpreted can be in the eyes of the beholder. Of course, it is extremely important to go beyond the quite specific experimental arrangement and the statistical results and to say something more general. For example, we usually do not wish to talk about how we operationalized the independent variable the measures (e.g., specific questionnaires, direct observation) but rather the constructs they reflect. For example, we would not want as our conclusion from the study, "Asking people to hold their breath and give a speech while someone is shouting in their face led to higher scores on the Lipshitz questionnaire of anxiety." We are looking for something more general such as "A stressful experience can increase anxiety." In other words, at some point in the study, we want to talk about the concepts our operational definitions (procedures and measures) represent and any implications for theory. Consequently, going beyond the data is desirable and actually essential. Yet, it is easy to slip into something that goes beyond what the design allows. Let me highlight main instances in which such leaps are likely to occur in clinical psychological research.

15.1.1: Common Leaps in Language and Conceptualization of the Findings

> "**Highly" and "Almost" Significant Effects**: Typically, results are evaluated with tests of statistical significance. Many if not most journals also require that effect sizes or strength of relationship measure also be included. Both of these often are unwittingly misinterpreted in a way that misrepresents the findings or at least what can be said about them.

Consider statistical significance. If a finding is statistically significant and the p level is computed as $p < .0001$, as opposed to the more modest, $p < .05$, we as investigators often are thrilled. (Remember that the lower the p value the stronger the support is not how null hypothesis testing works.) After all, we only needed to break the .05 barrier but look how great we did. Now we may refer to the results as "highly significant," which has no real statistical meaning or special role in null hypothesis testing. We thought the effect we were testing would be significant but never "this significant—wow—this was really significant." Anyone who shows joy in looking at data should be praised. After the praise is over, let us gently confront the problem.

Null hypothesis testing is based on demonstrating findings that are significant (meet the conventional level of $p < .05$ or .01) or not significant. Adding adjectives and words of enthusiasm ("highly, very, quite") is a lovely expression of joy but misunderstands the underpinnings of null hypothesis testing. In other words, "highly significant," in relation to null hypothesis and statistical testing is a mis- or overinterpretation. In other contexts (e.g., loving and romantic relationships), it might make sense to use "significant other" (I like the person a lot) or "highly significant" other (I may even love the person), but beyond that try to refrain.

> Apart from having no meaning in null hypothesis, saying "highly significant" or any of its partners advertises that one does not understand how this all works.

Of course it works in the other direction where $p < .06$, .07, and anything above $p = .05$ is not statistically significant. Many of us cannot bear to say that, even though we know, so we add "approached" to significance. (And in a new, budding, and possibly fine relationship, we might say things are "approaching significance.") Concepts and words of the same type are saying there was a "tendency" or "trend." Each of these misinterprets the quantitative finding. There is a very special term that covers each of these—you guessed it—"not statistically significant." We have all sinned on this, but if you engage in null hypothesis testing, your finding after a statistical test is significant or not. I am just the messenger here—one reason data evaluation has moved to adding effect size measures is precisely because of the limits of binary thinking and the vulnerability of "not significant" to low power. If you are reading a research report and your attention starts to wander, amuse yourself by finding how the investigator referred to almost but not quite significant effects. Again, the term should be "not significant" but you will find something else. There are alternatives to this binary decision making of yes or no and the frustrating one when we want to say almost yes ($p < .06$). For example, I have mentioned Bayesian analyses as one family of alternatives. Yet, if we as investigators play the null hypothesis statistical testing game, it is important to remember the rules.

15.1.2: Meaning Changes of Innocent Words and One Variable "Predicts" Another

Careful about Meaning Changes of Innocent Words: The word "significance" is misinterpreted in another way, beyond adding amusing adjectives such as "highly," "approached," or "almost-sort-of-please God." It is very easy to move from the concept and words "statistically significant," which of course has a very restricted meaning, to the word "significant," which has a very general meaning in everyday language.

Significance in this everyday use means:

- Important
- Noteworthy
- Meaningful

In our minds as we interpret the data, we often make this shift from significance in one sense (statistical) to another (everyday use).

A statistically significant effect can of course be important, but it can also be trivial (a comment my committee was fond of repeating during my dissertation orals).

> A statistically significant finding might mean important if the study is a proof of concept (e.g., identifying some new relation among constructs in psychology, showing a new particle in physics) but more often than not statistically significant has no necessary connection to significant or important as these words are used in everyday life.

If one were to do a word "search" of published articles, one can find "significant" used in these two different senses.

As much confusion and overinterpretation are accorded effect size (ES) as well. ES has conventional levels that are like many T-shirt sizes—small, medium, and large. ES is a statistical metric and gives magnitude strength of relations between variables. However, a finding, especially if a large ES, often is confused with practical importance or clinical significance. ES has no necessary relation at all to anything important in the lives of the participants, but is an important metric for other reasons.

In your own work, be cautious in using "significance" and "importance" if you are making reference back to some data analyses or reading a study. There are other issues like this we have already discussed. The terms "practical value" and "clinical significance" often are confused with effect size. Statistical significance and effect size have no necessary connection to these other terms and more often than not use of those terms go way beyond the data. Here can be a case where describing a finding might include hidden interpretations because terms are used that in fact are way beyond what can be said from the data analyses.

One Variable "Predicts" Another—Yes and No: Another common language/concept leap in a study is one in which several variables are evaluated at a given point in time (cross-sectional study), and the investigator uses statistical analyses in which the word "prediction" comes up.

For example, the investigator may have a conceptual model that notes certain cognitions and personality characteristics underlie depression. A sample of subjects is assessed, and any one of several analyses (e.g., structural equation modeling, multiple regression, discriminant analysis, logistic regression, to mention a few) is used to predict depression. These and other analyses classify variables as predictors (e.g., cognitions and personality characteristics, toss in ethnic group, sex, and socioeconomic standing perhaps) and the criterion or outcome (depression).

> The word "prediction" may emerge in the output from the statistical analyses. And the results, exactly as you expected, showed that a few cognitions and personality characteristics were statistically significant in "predicting" depression. In the context of the data analysis, prediction only means that the measures are correlated.

The investigator chose cognitions and personality characteristics as "predictors" of depression. But, all the measures were obtained in the study at the same point in time (time 1, because that is what a cross-sectional study means), and hence there only is correlation. Correlation can be informative, but that is not the point.

As the investigator moves to the interpretation of the results, a little leap is often made in the use of the word "prediction."

> That is, the investigator may use the word to imply a time line and to note that some cognitions and personality characteristic come together to *produce* depression. That is, we have subtly moved from correlate (all that really was shown in the data analysis) to risk factor (antecedent) for, or even a cause of, depression.

The problem is exacerbated by visual and graphical representations of findings (e.g., structural equation modeling, mediation analyses) in which there might be arrows that point in a direction, which further suggests directionality and prediction from one point to some next point. This gives an illusion of directionality and may even lead to one thinking about causality.[1]

Any time one sees the word "prediction" in a discussion of the results, it is important to be mindful of whether the design warrants the use in which a time line is implied. In a cross-sectional study in which all of the measures are administered at a particular point in time, one has to be especially cautious. As we read the write-up of the study, we may see nothing worth questioning (e.g., threats to validity) in the design. Then we may see the author's discussion and see whether she wanted to talk about "prediction" and "causal"

relations. Now we see the design differently. For those conclusions, the design was inadequate because no time line was allowed by retesting over time. Either the design has to change (too late for that) or the discussion has to be brought in line with it (never too late).

15.1.3: "Implications" in the Interpretation of Findings

> "**Implications" of My Findings**: A final illustration in the move from data analysis to discussion and interpretation of the findings has to do with the notion of implications. In a study, we discuss the implications of our results. These refer to how the specific findings of this study might have consequences well beyond the demonstration and the narrow or specific laboratory paradigm we set up to test our hypotheses.

Implications are critically important of course, and it would be futile if we bounced from one study to another that was of no help in drawing broader conclusions beyond the necessarily very narrow restrictions of that demonstration. So my comments here are not "against" discussing implications. Actually, the implications section of a discussion may stimulate new lines of work and make connections that the reader would not otherwise make. Rather, I mention the issue to continue the focus on instances where too much or too far can lead to overinterpretation of the findings. What is "too much or too far" has no clear definition (like the binary nature of statistical significance), but entertain a few considerations.

The problem is that "implications" is sometimes used as a ticket to enter any topic one wishes. In clinical psychology and related areas, this emerges most often in drawing implications about treatment and prevention from a study that did not really focus on interventions and might be argued as rather far removed. For example, we know that teen alcohol use and abuse is related to several factors, including whether one's parents drink alcohol in the home, conflict in the family, being in a home with a single parent, poor peer relations, and many other influences (e.g., Curcio, Mak, & George, 2012). One such influence is whether the family eats meals together on a regular basis. In a well-designed study of youth throughout the United States, the focus was on the onset of alcohol use. Youth (9–14 years of age) who had family meals together were much less likely to engage in alcohol use when evaluated 2–3 years later (Fisher, Miles, Austin, Camargo Jr, & Colditz, 2007). (I mentioned this study previously in the context of using single items, in this case one self-report item[!], to measure eating family meals together.) This was a prospective study, so we know the time line—meals together preceded onset of alcohol use. And we have a "predictor" in both senses of the word (statistical and "real") precisely because that time line was established. We have no idea of whether having family meals was related to many other characteristics (e.g., parent psychopathology, family history of alcohol use, antisocial behavior of the child) and all those are suitable topics for additional research. But an "implication" was drawn that one has to read carefully.

In moving to implications, the authors noted, "Having family dinner together on a regular basis may be a simple way for parents to reduce the chances that their children will initiate alcohol use" (Fisher et al., 2007, p. 7, online). To the authors' credit, they worded the statement with the subjunctive "may" to place appropriate tentativeness on the possibility. Yet for the present discussion and a more strict interpretation, there is no reason within this study to suggest that changing eating meals together would have any impact whatsoever on alcohol use. The move to suggesting an intervention arguably might be too far. We really do not know it was even eating family meals together (construct validity problem) because that was assessed with a single face valid item (not clear what was measured, eating meals? Social desirability? How much one likes one's parents?). Also, eating meals together could be a proxy (stand for) so many other variables, including a strong genetic and environmental "loading" for antisocial behaviors such as drug use, fighting, and stealing. And finally, maybe family meals were eaten together only among parents and children who could get along and were not suffering from clinical dysfunction. Family meals together could readily be confounded by other influences that cause eating or not eating together.

Did the authors go too far in their implications? No individual methodologist (e.g., me) is the arbiter of what implications can be stated that is at the Goldilocks (just right) level and no doubt I have sinned somewhere along the way. (I probably should not have spoken quite so much on the "intergalactic implications and relevance" of my dissertation results.) The eating meals together example went from a correlation (with an established time line) to possible cause (manipulate the correlate of eating together and the outcome of alcohol use might change). This is a broader issue of keeping conclusions (inferences) close to findings (actual results). Discussion of "implications" is a place where these might be mixed up. The issue is mentioned to be cautious in your own work and reading the works of others; be skeptical when you see some leaps and the occasional word "implications" that can be a flag.

15.1.4: Further Considerations regarding "Implications"

More generally, in the study of risk and protective factors, it is relatively common for investigators to move to the preventive implications of the findings. As you recall, risk and protective factors are antecedents that relate to some future outcome (e.g., drug use, psychiatric disorder, obsessive

love of methodology). Once a risk or protective factor is identified, investigators are wont to leap a little from what was shown to what we ought to do about it, i.e., an implication. Specifically, we say that interventions ought to target (focus on, change, redress) one the antecedents. When risk factors (eating meals together) in this way are malleable (e.g., could be changed by intervention) as opposed to more fixed or at least difficult to change right now (e.g., personal genome, culture of origin), we tend to move to intervention implications without appropriate qualifiers. Thus, eating meals together at time 1 relates to alcohol use at time 2—let us tell the world to eat meals together to reduce alcohol use. That is a nonsequitur. The reason is that a risk factor may not be causally related to the onset of the problem, and changing the risk factor may have absolutely no impact on changing the problem.

Consider a more concrete (and personal example that applies to a "friend" of mine) illustration. We know from many studies that among men, being short and bald are risk factors for a heart attack—and we know each comes before having a heart attack because of prospective studies and the logical fact that few people have a heart attack only then to be hit with short height and baldness (e.g., Paajanen, Oksala, Kuukasjärvi, & Karhunen, 2010; Yamada, Hara, Umematsu, & Kadowaki, 2013). So we have some risk factors (antecedent correlates); what are the implications? Well, I or rather my "friend" may have gone too far perhaps. To reduce risk, he wears 4-inch spike heels so that being short is no longer relevant; also he now wears the thickest imaginable toupee, so being bald is off the table. Now with changes in these two risk factors, he feels a little safer.

> **The lesson:** We do not focus on something for purposes of intervention just because it comes before the problem; the something could be a proxy for the variable that really makes a difference and even if the variable (family meals, baldness) is perfectly accurate that still does not mean intervening will help.
>
> This has huge "implications" in serious contexts. Poverty precedes high rates of many diseases and early death.

Will changing poverty alter the outcome?

Unlikely, poverty ought to be actively ameliorated for a variety of important reasons, but lack of funds per se is not likely to be causally related to these very sad outcomes. Making people unpoor or less poor (e.g., providing higher wages and more money) may not likely to lead to healthful habits (e.g., reduced cigarette smoking, higher child vaccination rates, access to regular medical care, increased consumption of healthful foods, increased exercise, reduced levels of obesity), which are related to poverty and probably closer to factors that influence the outcomes. Active efforts are needed to ensure access to the benefits and services that promote health and in general reducing disparities in care.

In general, authors often feel compelled to address applied implications (e.g., for education, treatment, prevention, physical or mental health). In many cases, there are external pressures (e.g., grant proposals in which the impact of the work sometimes needs to be far reaching, requirements journal editors invoke on authors to move well beyond the findings) to convey possible implications and broad relevance of the findings. Indeed, funding agencies that provide grants often are under the same pressure in arguing for important "implication," because legislators (state, federal) ask why would we want to fund a project about . . . (select any of these: zebra fish, brain processes in the social behavior of mice, sexual attraction among college students, and so on). So making connections of one's research project to something that might actually make a difference somewhere on the planet is important and often required.

My recommendation—be mindful of the stretch in discussing what leaps to make from the findings. Findings may be important in their own right because they elaborate basic issues about affect, cognition, behavior, clinical disorders, adaptive functioning, and so on.

> We want implications for theory and understanding as much as for application, and these two former types of implication receive less attention. What are the implications of the study for theory about the phenomena you are studying and from that what should be the next study or few studies that are really needed to advance the area.

These are implications that are to be encouraged. If one goes to application (treatment, prevention, rehabilitation, day care or educational practices), merely be circumspect, qualified, and super careful. There is a natural tension from what you might want to say (e.g., world peace will come from eating meals together), what you are being asked to do ("why should we fund or publish this?" from funding agencies, journal editors), and what can be said based on the design of the study ("this is a very interesting finding in light of these theoretical considerations, prior research, and 'may' be a basis for intervening"). Also, if you really believe changing a risk factor will make a difference, by all means do the study to show that is true.

15.1.5: More Data Analyses Can Enhance Data Interpretation

In most research, the investigator is searching for overall group differences. That is, some comparison is made (e.g., experimental manipulation vs. a control condition or various pre-existing groups such as patients with a diagnosis of bipolar disorder or major depression). The investigator is looking for a main effect (overall effect) of the two (or more) conditions. Although there may be an overall effect, it is unlikely that the overall effect applies to everyone. That is, not all people in the experimental condition changed in the predicted

direction and not all people in the other or control condition remained the same. There is variation within a group and condition and this is common. Sometimes we expect or hope for something more or different. For example, in treatment studies using evidence-based treatments (e.g., psychosocial interventions, medications, surgery), not everyone responds even when the treatment was administered as planned. And of course, among those who did respond there is variation in the degree of responding.

For example, one study evaluated the effectiveness of medication in treating depression and whether there were differences in responding among European-American and African-American patients (Lesser et al., 2010). Both ethnic groups responded equally well. Overall approximately 50% of patients responded to the treatment. This is not an aberrant finding, and other research has found that with a number of medications and forms of psychotherapy (e.g., interpersonal psychotherapy, cognitive behavior therapy), 25% to 50% of the clients may not respond to treatment, despite an overall group effect (American Psychiatric Association, 2000). The 50% rate begs for further evaluation and analyses:

- Who responds and who does not respond and why?
- What are the critical variables that might identify responders and nonresponders?
- How do these variables operate?
- Precisely what might be done to increase response rates or to redirect patients to improved treatments?

We have a main effect (medication worked). But, can we do more to understand the variation in responding?

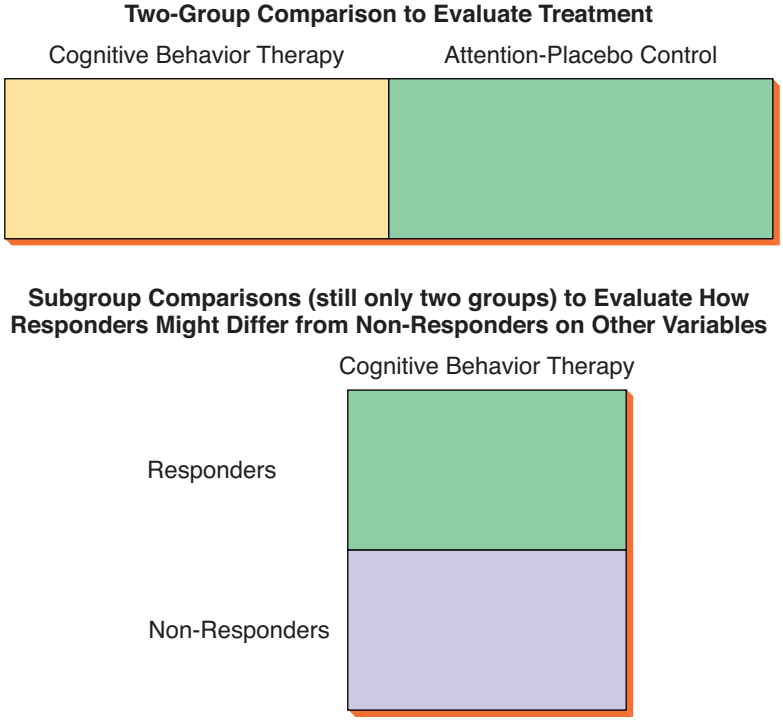
> **Exploring Treatment Moderators:** Let us say, we are interested in comparing cognitive behavior therapy (CBT) versus an attention-placebo condition for the treatment of depression. At posttreatment we compare the groups on several measures of depression.

There are two cells (groups) in the study, as illustrated by Figure 15.1 (top panel). It might be useful, informative, and helpful as a guide for subsequent research to explore this more fully, more about these groups that might elaborate who responds to the treatment.

In the lower panel, the CBT group is divided into two subgroups based on whether subjects responded well to treatment (responders) or did not (non-responders). The data can be explored to identify what characteristics differentiate these groups. Likely candidates for analysis in a clinical sample might be severity of clinical dysfunction at pretreatment, presence of comorbid disorders, or of course variables that are informed specifically by the theory underlying the therapy or epidemiological findings on the factors that contribute to onset, course, and prognosis of the clinical problem. These are all post-hoc analyses rather

Figure 15.1: Hypothetical Example: Analyses of Treatment Effects in a Two-Group Study

A two-group study might be analyzed to make the direct comparison of treatment (upper panel) or to examine influences of possible moderators, i.e., factors on which treatment effects may depend (lower panel).



than predicted and one must be especially vigilant about limitations (e.g., power may be weak, many tests might be done to maximize chance findings). At the same time, not to look at the data has its own drawbacks since multiple clinical studies are difficult to do and the informational yield for testing and generating hypotheses ought to be maximized.

For example, consider that we could classify all CBT participants on the basis of whether they improved (e.g., reduced in their symptom scores by some arbitrary unit such as >1.00 of a standard deviation on a key outcome measure or set of measures that was used to form a single index). Let us say, now we have two groups but they are different groups from before (bottom panel of Figure 15.1). We have omitted the placebo group (but easily could include them as well) and are looking at who improved and who did not within the CBT group.

We might analyze responders and nonresponders to see if there are any differences at pretreatment. The purpose of these analyses is to generate hypotheses about what might be pertinent, helpful, or relevant to responsiveness to treatment. Because all of this is post hoc and exploratory, many statistical analyses are likely to be conducted. Special care is needed in interpreting the findings. That is, mining the data may have greater opportunities for chance effects if for no other reason than multiple tests are conducted and experiment-wise error rates are not controlled. We are using these analyses to *generate* hypotheses. We may even want a lenient p level (e.g., $p < .10$ or $< .20$) just to identify tentatively what might be related to responsiveness. Exploratory data analyses can be lenient, but data interpretation has to be more rigid. The explorations are tentative; just be cautious in over interpreting what you find. The value of the analyses is to ponder how any differences might explain the findings and to use that information as a basis for another study. That is, we are *not* doing this exploration to detect statistically significant differences that we pluck out and report as if we expected this all of the time (see Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). The data evaluation here has nothing to do with publication or writing up a report but rather with understanding what might be operating—we are looking for leads. We may have an idea who responded and who did not (theory), but we can also explore the data to generate that theory.

15.1.6: Another Example of More Data Analyses Enhancing Data Interpretation

Looking for subgroups in this way can be anywhere from uninspired to interesting and provocative depending on what the investigator does and how he or she makes use of the information in later research. In an example from my own work (perhaps toward the uninspired end of the continuum), we have worked on evidence-based treatments for children with aggressive and antisocial behavior (see Kazdin, 2010). In some of the studies, we have compared children who completed and therefore received the treatment with those who drop out very early and therefore receive little or no treatment. As one might expect, those who complete treatment do much better those who do not. (When we first found this, I sent the Nobel Prize Committee my correct address just in case they were not contacting me because they had my old, incorrect address.) Ok. Ok, so this finding is not so special. However, the subgroups become rather interesting—among those who complete treatment, many (~78%) make large changes and get a lot better but others make smaller changes or do not get better. Among those who drop out very early (first few sessions), some (~34%) get a lot better, but many do not.

> What distinguishes all of these children, and can this information be used in any way to inform treatment?

Among the factors, parental stress plays some role because it was related to both who drops out of treatment and who profits from treatment. But the methodological point is the one to emphasize, namely, exploring the data can generate insights that might be used to improve treatment. There are many variables that differentiate who do and who do not respond but let me mention one that we know a little more about.

Parental stress in the home influences child treatment outcome. Children whose parents are more stressed are less likely to profit from treatment. The treatment we were evaluating was parent management training (PMT) (Kazdin, 2005), and hence parental stress might directly impede how well parents can benefit from the training. Highly stressed parents still make gains and their children improve, but not to the same degree as those from parents who are less stressed. As I mentioned, our post-hoc, exploratory analyses found that pretreatment stress influences (moderates) child improvement at the end of treatment. Perhaps children in homes with greater stress are more severely impaired, and it is severity of child dysfunction that is the critical variable. Indeed, parents might be more stressed in part because they are dealing with such children on a daily basis. Yet, there were reasons to believe from other research on parenting, family life, and child behavior that stress may interfere with benefitting from treatment (Deater-Deckard, 2004). Does stress play role in child treatment outcome beyond merely being a predictor? There are a few ways to answer the question. One would be to repeat the study and include a more extensive battery of measures to see if it is stress or one of the many variables (e.g., poverty, single-parent homes, conflict with prior partner) with which it is associated. Another way would be to

directly intervene and change parental stress as part of treatment. We elected the latter strategy—it gets to the question directly—if we change parental stress do we change treatment outcome?

We conducted another trial providing our usual parenting treatment to some and providing that same treatment to others but with an added stress management component (random assignment of course) (Kazdin & Whitley, 2003). The stress management focused on problem-solving strategies to cope with stress and then actually assignments in everyday life to address the stress. Parents with the stress management intervention showed predictable reductions in stress and reductions in stress that were greater than parents who did not receive that treatment component. Thus, the stress intervention in fact altered levels of parental stress.

Did reducing stress have any impact?

Children in both groups improved, but those in the group whose parents received the stress intervention improved more. Our post hoc analyses in the prior study provided useful leads about what we might focus on and alter in the next study. All of this answered one little question and raised many others (e.g., stress does not operate in the same way for everyone, what is it about stress that makes parents respond less well, and so on).

The overall point—explore data for potential leads. Try to find who did respond to some experimental manipulation or intervention really well or not at all. Perhaps also look at the control conditions—who responded there even though they were not expected to. If possible interview participants and experimenters about specific cases. We are doing the equivalent of scientific brainstorming to generate leads and any source that could help make the next project all that better and more informed.

15.1.7: Searching for Moderators or Statistical Interactions

Predicting Moderators: The search for subgroups or factors that influence who responds to an experimental manipulation or intervention is the search for *moderators* or for statistical interactions, rather than main effects.

The interactions reflect the fact that the impact of one variable is not equal across another condition (e.g., sex, severity of dysfunction) but rather varies systematically as a function of that other condition. The search for these other conditions in a post hoc way as noted previously is useful, particularly as a guide for subsequent studies.

If at all possible, it is especially useful to predict interactions among variables.

Predictions about the interactions of independent variables often reflect greater understanding of how the independent variables operate than do predictions about main effects. Interactions begin to delineate the boundary conditions for a particular effect or experimental variable.

As the boundary conditions for the effects of a given variable are drawn, a more complete understanding is available than from the demonstration of a straightforward main effect. Let us say you are conducting a study on physical attractiveness on whether college students would want to date or be with a new person they meet. You set up an experiment where students view different faces and have already from pilot work or other studies have photos of cases (male, female) that differ on attractiveness. You believe that another variable (trustworthiness) will moderate subjects' ratings so that attractive photos will not be rated as people to be with by the subjects if those attractive people in the photos are low in how trustworthy they are. You accompany photos with a description to convey that the individual can be trusted (e.g., with secrets or someone else's money) or cannot be trusted. Perhaps we find in this experiment that physically attractive photos are rated as better to be with on a date (main effect) but that this effect is muted, small, and not so strong if the photos are of people who are not very trustworthy. That is, attractiveness indeed is influenced (moderated) by trustworthiness. This is an interesting finding by showing that attractiveness and its evaluation depend on another variable.

In general, prediction of moderator effects is an excellent base for research because it can contribute to a deeper understanding of mechanisms and processes that explain how variables work and why their effects vary under different conditions.

Consequently, findings are viewed in the context of how they draw from and contribute to theory and understanding of process. Statistical interactions can provide important leads by identifying what factors are important and by prompting considerations of how these factors might operate, i.e., why they are important.

15.1.8: General Comments

Research would be much simpler if variables operating in the world were restricted to main effects. Results of experiments could be accepted or rejected more easily if a given variable were always shown to have either an effect or no effect. Because variables often do interact with each other, it is difficult to interpret the results of a single experiment. If a variable has no effect and groups were not different in tests of statistical significance, it is always possible that it would have an effect if some other condition of the experiment were altered. That is, the variable may produce no effect for certain subjects, experimenters, or other specific

conditions but later produce great effects when any of these other conditions is altered. Often in research, we question the external validity (generality) of the findings obtained in a study. Equally, we can question the generality of the findings when an effect is not obtained, i.e., whether the variable would have no impact if tested under other conditions. One of the tasks of research is to identify the circumstances under which some relation would and would not be obtained.

There is a related but more subtle issue. Two groups may be no different in a given study. You have done all the statistical tests possible (a problem we discuss later) and nothing "came out." Quite possibly, buried within that the groups are subgroups that might consist of a few or many people who in fact responded well. That is, some moderator or interaction may be hidden. Yes, of course, the responses of a subgroup one identifies post hoc could be due to chance, statistical regression, or be answers to your endless prayers for something systematic in the data? It could also be something important or interesting.

> One reason to do exploratory analyses is to understand one's own data and to identify if there are patterns, interesting exceptions, and possible subgroups. The goal is always to understand one's data.

When other people do these analyses, the pejorative term "fishing expedition" is often used. When you and I do these analyses, they are called "exploratory data analyses." There are some real differences in the terms sometimes. In my exploratory analyses, I am not looking for "statistical significance" to pluck out a chance finding and salvage a study that does not otherwise seem publishable. (I already did that for years and gave up with my dissertation and found "nada.") Are there possible patterns hidden in the data that might prompt further thought, theory, and another study? The value of one study often is how it informs the next study.

In one sense, variables studied by psychologists virtually always can be considered to interact with other variables, rather than to operate as individual main effects. Even if no interactions emerge in the analyses of the results, other conditions than those included in the design or data analyses may influence the pattern of results. The conditions of the experiment that are held constant may reflect a narrow set of circumstances under which the variable produces a statistically significant effect. The effect might not be produced as these conditions change. Obviously, the effect of variable X on Y may depend on age of the subjects (infants vs. adults) or species (e.g., primates vs. nonprimates). Few, if any, results obtained by psychologists would be replicated across all possible variations in conditions that could be studied; that is, there are implicit interactions among the conditions studied.

15.2: Negative Results or No-Difference Findings

15.2 Explain why no-difference findings occur in scientific research using the five reasons

In most investigations, the presence of an effect is decided on the basis of whether the null hypothesis is rejected. The null hypothesis states that the experimental conditions will not differ, i.e., that the independent variable will have no effect. Typically, rejection of this hypothesis is regarded as a "positive" or favorable result, whereas failure to reject this hypothesis is regarded as a "negative" result. Advances in research usually are conceived as a result of rejecting the null hypothesis. Recall that the null hypothesis significance testing (NHST) model is *one* model of doing science and there are other options (e.g., Bayesian analyses, model building) that operate differently. That said, the null hypothesis testing model remains by far and away the most dominant model in psychology and sciences more generally. The dominant model is the context for this discussion.

> As researchers know all too well, many investigations do not yield statistically significant findings or any other evidence that the independent variable influenced the participants. The term *"negative results" has come to mean that there were no statistically significant differences between groups that received different conditions or that the result did not come out the way the investigator had hoped or anticipated.*

Usually, the term is restricted to the finding that groups did not differ, leading to acceptance of the null hypothesis.

> The presence of a statistically significant difference between groups, i.e., a "positive result," often is a criterion—indeed a major criterion—for deciding whether a study has merit and warrants publication. This is referred to as a *publication bias, namely, favoring publication of studies that find differences.*

The bias is of enormous concern in scientific publication generally (e.g., Joober, Schmitz, Annable, & Boksa, 2012; Johnson & Dickersin, 2007). The bias is real: studies with "positive" or statistically significant results are more likely to be published (Dwan et al., 2008). The publication bias is not a minor annoyance. It is quite possible that many published research findings are false (i.e., due to chance) and would not be replicated (Ioannidis, 2005), a topic we will discuss later in the chapter. But leaving aside "true or false" for a moment, it is clear that the bias distorts what we might conclude.

For example, a comparison of published studies evaluating medications for the treatment of depression with reports of those same studies to the Food and Drug Administration allowed one to examine the extent to

which positive and negative results were published (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). When a medication was effective (97% for positive result), the finding was much more likely to be published than when the medication did not show an effect (12% negative result). Conclusions about the effect of an antidepressant medication clearly are biased by the very selective publication of the results of clinical trials. The result also conveys that occasionally, "negative results" are reported in the published literature. Yet even then, they are more likely to be published in less prestigious journals and therefore probably less visible among the scientific community (Littner, Mimouni, Dollberg, & Mandel, 2005). Clearly the pressure is not to report one's data, but to find some differences.[2]

The search for group differences so that the results will be publishable may encourage sacrifice of methodological standards, the possibility of inadvertent bias, or outright dissimulation. (This pressure contributes to fraud and lapses in scientific integrity.) Poor methodology and sources of experimental bias are more likely to be overlooked when a predicted or plausible finding of significant differences is obtained. The implicit view is that group differences demonstrate that, whatever failings of the experiment, they were not sufficient to cancel the effects of the independent variable. In contrast, negative results often imply that the independent variable was weak or that the study was poorly designed or conducted. One is almost "blamed" for how things came out, when our goal is primarily to find out how things are and why.

The value of a study can be assessed as a function of its conceptualization and methodological adequacy, rather than, or at least in addition to, whether differences are found. The conceptualization refers to the importance of the question, the theoretical underpinnings (if applicable), and how well thought out the question is, as described in the report of the study. The methodology refers to all those facts that reflect threats to validity and sources of artifact and bias. Conceptualization and methodology of an investigation bear no necessary relation to the outcome or statistical significance of the findings. A sloppy, ill-conceived, and horribly uncontrolled study can lead to systematic group differences (as my dissertation committee was all-too-eager to point out).

> As investigators, we wish to proceed with the best or strongest available design and greatest methodological care so that the results of the study, whatever their pattern, will be interpretable.

Assuming that the question addressed in the investigation is important to begin with, methodological adequacy of the design, rather than pattern of the results, ought to be the main criterion for evaluating the study. This latter point has been advanced for decades and continues to be (e.g., Greenwald, 1975; Kupfersmid, 1988; Lykken, 1968; Pfeffer & Olsen, 2002), which means it has yet to be widely adopted. The difficulty in judging the value of any study is that neither the importance of a finding nor methodological adequacy is invariably agreed on by those who do the judging. Rejection of the null hypothesis and statistical significance are overly relied upon because they present relatively simple bases for evaluating research.

15.2.1: Ambiguity of Negative Results

The absence of group differences in an experiment is not usually met with enthusiasm by the investigator or by the reviewers and journal editor who may be considering the manuscript for possible publication. This reaction derives from the ambiguity usually associated with negative results. The reason for or basis of "no-difference" findings usually cannot be identified in the experiment. The most straightforward reason for accepting the null hypothesis is that there is in fact no relation between the independent and dependent variables. There are, however, many other explanations for a "no-difference" finding. Table 15.1 lists some of the main reasons and they are worth highlighting briefly.

Table 15.1: Some of the Obvious Reasons for No-Difference Findings

Obvious Reasons for No-Difference Findings
1. There are no or very small differences in the population, i.e., the no-difference finding reflects the true state of affairs. (Power was low and probably too weak to detect a difference.)
2. The investigator could not duplicate or carry out the manipulation or intervention or the manipulation was not carried out as intended (e.g., diffusion of conditions, poor adherence of experimenters or therapists, groups were not different on a manipulation check)
3. Levels of the independent variable (e.g., low, medium, high) were not optimal or did not provide a strong test
4. The assessment devices may not be sufficiently sensitive to reflect real differences (e.g., due to highly variable performance of the sample, ceiling or floor effects, and considerable error in the measures)
5. Excessive uncontrolled "error" variability (e.g., heterogeneity of subjects, loose procedures for implementing the study)

Each of the above reasons is discussed in detail in the following sections.

First, one possibility is that the study had insufficient statistical power to detect a difference that in fact was there. There are other statistical issues that could easily explain the absence of differences, such as whether the assumptions of the data analyses were met, whether adjustments (to control for experiment-wise error rates)

were or were not made in the number of tests, and whether outliers (subjects with quite extreme scores) were or were not included in the data analysis, to mention a few. Yet, weak statistical power for detecting small-to-medium effects continues to be the rule in psychological experiments, so this reason alone could explain many of the no-difference findings. A recent analysis of power of psychological studies suggested .35 is the average power to detect an effect if there is one (Bakker, van Dijk, A., & Wicherts, 2012). Obviously, this is well below the recommended power of .80 discussed previously and why emphasis was accorded the importance of increasing power. We do not have to go much further in looking for reasons—weak power is so pervasive.

Second, no differences could easily result from the fact that the experimental manipulation was not carried out as intended. There are separate variations of this. For starters, it is possible that no differences between conditions occurred because of a *diffusion of treatment*. This was discussed as a threat to internal validity and can mean that not everyone in the experimental or treatment group or in the control group received the condition that was planned. It could mean that some of the people assigned to treatment really did not get the intervention or some people in the control group somehow received the treatment. This will blend the two groups or conditions and operate to make them less or no different.

Apart from diffusion of treatment, lapses in treatment integrity or treatment fidelity could explain no differences. Perhaps the manipulation treatment was not delivered as intended. Here there is no diffusion necessarily. That is, everyone in the treatment group received the condition to which they were assigned. Similarly, everyone in the control group (e.g., treatment as usual or no treatment) received what they were supposed to receive. No treatment diffusion. Yet, treatment delivery was not monitored or assessed and was delivered partially, inconsistently, weakly, poorly, or some other such adjective. This would have two effects that could lead to no differences. The inconsistency in delivering treatment would introduce extra variability (error) and reduce the likelihood of obtaining statistical significance. Also, some participants received poor doses of the treatment and that too would diminish treatment differences when compared with a test of statistical significance. A check on the manipulation may show that a critical facet of the manipulation was not provided.

To make this more concrete, think of a treatment study of people with headaches assigned to the pain reliever or no-treatment control condition. The treatment (e.g., aspirin or the equivalent) dose may be two pills, but some people only get a quarter of just one pill (1/8th of the dose). When treatment is not done correctly or as intended or given in the full dose, the lack of differences between or among conditions is much more likely than it would otherwise be and could account for "negative results." Indeed, we know from many studies that whether and the extent to which treatment is delivered correctly influences treatment outcome (e.g., Huey, Henggeler, Brondino, & Pickrel, 2000; Strauss et al., 2012).

We often underestimate the more gross point, namely, can the investigator carry out the experimental condition or intervention at all? For example, assume I invent a new therapy (such as my latest discovery) called *Mindlessness* (one of my more clever interventions for people who do not respond to Mindfulness). Let us say I show in a controlled trial that this treatment is effective. Now you come along and fail to replicate my results. Ought I to believe this or any other failure to replicate? Only my research group and I are capable of carrying out the treatment, and your test may not be a good one because you simply do not do the procedure correctly. Negative findings can result when the investigator does not know how to implement the intervention or fundamental procedures to carry out the study. More seriously, this comes up on other contexts.

For example, within psychology, Ivan Pavlov is recognized to be an extraordinary scientist, as reflected by his earning a Nobel Prize (in 1904) for his work on digestion. He also devised a technique to isolate a portion of the stomach to assess gastric secretions. The surgical techniques to accomplish this were quite difficult, and several dogs died in the process of perfecting the surgery (Cuny, 1965). Of course, Pavlov's work on conditioning has been greatly replicated and extended, but the procedures initially were sufficiently difficult that it would be understandable if many could not even carry out key aspects of the procedures or did so in ways that did not permit the meticulous assessment (individual drops of saliva) that Pavlov achieved. In clinical psychology, we worry about treatment integrity, but there are analogs in any area where a task is required of the subject or special talent, skill, or training is required of the experimenter. Poorly implementing a difficult experimental procedure could readily lead to a no-difference finding.

Third, the absence of group differences may also result from the levels of the independent variable selected for the study. Whether or not differences are obtained may be completely determined by what levels were selected. For example, whether and the extent to which there is a relation between the amount of the intervention or experimental manipulation and change on the dependent measure. When we carry out a psychological intervention, we rarely have an idea of how to make it really strong as opposed to really weak. Theory can be very helpful because it may propose what facets are the main reason that change will occur and by that influence what to optimize and maximize in the study.

Selecting the level or strength of an intervention is more easily worked on when medications are studied. Prior to intervention work, studies can evaluate dose in quantitative terms (e.g., how much is given, what the blood levels are, whether and when side effects emerge). Psychological manipulations in the laboratory and psychosocial interventions in clinical settings are proceeding a little more blindly. A given hypothesis might not be supported, and the experimental conditions may show "no differences," because the "dose" or equivalent of the experimental manipulation was not that strong. In pilot work or within the study itself, different variations of the experimental manipulation might be tested. In any case, one reason for a no-difference finding is due to the specific level(s) of the manipulation what was selected. A different "dose" or way of implementing that manipulation might have yielded different findings. A challenge is to make the manipulation strong, but more or stronger is not always better because all relations are not linear.

Fourth, the assessment devices used to show the impact of the experimental manipulation may not be sufficiently sensitive to reflect real differences. This could be the result of highly variable performance of the sample, ceiling or floor effects, and considerable error in the measures. Home-made scales (invented for a study and only have face validity in their behalf) that include one or a few items would be an example where the intended effect was not obtained. We are left with the ambiguity of what was being measured and whether the scale(s) could reflect change. In addition, the dependent measures may not be the most appropriate for detecting a relation of interest.

Fifth, any factor in the experiment that operates to increase within-subject variability also may increase the likelihood of a no-difference finding. As we discussed in relation to data-evaluation validity and power, the magnitude of an effect can be measured in terms of effect size, or the difference between means divided by the standard deviation. A given difference between means is lower in the effect size that is generated as the standard deviation (denominator) increases. The sensitivity of the experiment in detecting a difference can be reduced by allowing uncontrolled sources of variation or "noise" into the experiment. Allowing such factors to vary as the adequacy of training of different experimenters and methods of delivering instructions can increase the variance within a group (experimental or control) and reduce the likelihood of finding group differences. Indeed, negative results have on occasion been implicitly used to infer that the investigator may be incompetent for not controlling the situation well enough to obtain group differences (reflections from my dissertation committee again).

For these and other reasons I have covered, one can easily see the ambivalence in evaluating negative results. So many reasons might explain the finding of no difference that many believe there is unresolvable ambiguity with what we can learn. This is understandable but arguably uninformed. There are truly excellent, well-designed studies that find no difference. We want to know about them, so other investigators will not continually pursue the same leads without knowing that these have been tried endlessly without success. Also, in the case of life and death issues (e.g., cancer treatment) and impairment and quality of life issues (e.g., treating posttraumatic stress disorder [PTSD], coping with a severe disability), we want to know what seemed to be well tested but did not work. More on this topic of when negative results are interpretable and important is provided next.

15.3: Why Negative Results Are Useful

15.3 Analyze the utility of the negative result in statistics

There are numerous situations in which negative results are very informative, interpretable, and thus extremely useful. These situations and circumstances are described in detail below.

15.3.1: When Negative Results Are Interpretable

The absence of group differences is routinely dismissed as ambiguous and often is given much less attention than it should receive. There are numerous situations in which negative results are very informative and interpretable and of course many situations in which "positive" results are not very informative, ambiguous, and misinterpreted. Negative effects are interpretable under a number of conditions.

First, in the context of a program of research negative results can be very informative.

A program of research refers to a series of studies conducted by an investigator or group of investigators. The studies usually bear great similarity to each other along such dimensions as the independent variables, subjects, and measures. Presumably several studies would have produced some group differences (otherwise it would be a masochistic rather than programmatic series of studies). The demonstration of group differences in some of the studies means that the experimental procedures are sensitive to the effects of the experimental manipulation or intervention. Thus, one can usually rule out the problem that the experiments are conducted poorly or that the methodology is too insensitive to detect group differences, even though these explanations may be true of one of the experiments. Several of the reasons for "negative results,"

mentioned in Table 15.1, become slightly less plausible than they otherwise might be precisely because the program of research has established itself in terms of demonstrating group differences across repeated studies. A new study showing no differences for a related variable can be viewed with greater confidence than would be the case in an isolated study. Thus, the results are likely to be more readily interpretable where there is a string of demonstrations attesting to the general paradigm because of its use in prior studies.

> **Second**, negative results are also informative when the results are replicated (repeated) across many different investigators.

A problem in the psychological literature and perhaps in other research areas as well is that once a relation is reported, it is extremely difficult to qualify or refute with subsequent research. If negative results accumulate across several studies, however, they strongly suggest that the original study resulted either from very special circumstances or possibly through various artifacts. Failures to replicate do not invariably influence how the field interprets or views the status of the original finding.

A classic example is the well-known study of Little Albert, an 11-month boy, whose story is taught to almost every undergraduate. In this demonstration, a loud noise (and a startle reaction) was paired with an object (white rat) that had not previously evoked a reaction from Albert. After several pairings, presentation of the rat alone led to the startle response (Watson & Rayner, 1920). This study has been extremely influential and is cited often in texts (e.g., like this one), even though several failures to replicate are well documented (see Kazdin, 1978). The replication failures did not challenge the finding that fears seemingly could be conditioned.

> **Third**, negative results are informative when the study shows the conditions under which the results are and are not obtained.

The way in which this is easily achieved is through a factorial design that permits assessment of an interaction. An interaction between the different factors indicates that the effect of one variable depends upon the level of another variable. An interaction may be reflected in the finding that there are no differences between groups for some levels of the variables but statistically significant differences between groups for a different level of the variables; that is, negative results occur in only some of the experimental conditions. For example, one of the goals of personalized medicine is to find characteristics that can be used to decide what treatment to provide. A treatment (e.g., medication) may or may not be effective depending on some other variable (e.g., genetic makeup). In the language of this chapter, negative or positive results will be influenced by that other variable (e.g., genes) and we would want to know that.

A related way in which a no-difference finding is informative is in relation to the pattern of results across multiple measures. A no-difference finding may be evident on some measures but not others. A curiosity of evidence-based psychological treatments is that in many studies, dependent variables used to evaluate the impact of treatment in fact show no difference (De Los Reyes & Kazdin, 2006). These measures are usually not mentioned or emphasized—the measures that did show predicted effect are used to support the hypotheses. Here the issue is not negative results of a study per se, because some measures did show the positive results. Rather, the issue is how information is used and reported.

As a general rule, when an investigator can show within a single study that a particular relation does and does not hold depending on another variable or set of circumstances (another independent variable) or does not hold across measures (e.g., dependent variables), the study is likely to be particularly informative. That other variable is of course called a moderator. Such studies often provide a fine-grained analysis of the phenomenon.

15.3.2: When Negative Results Are Important

Noting that negative results are interpretable suggests that a no-difference finding is salvageable. Actually, in a variety of circumstances finding no differences may be extremely important. In clinical and applied research, no-difference findings may be especially important in the context of possible harm, side effects, or costs of alternative procedures or interventions. As citizens and consumers of research, perhaps more than in our role as investigators, we care very much about and are actually rooting for many "negative effects," i.e., no-difference findings, across many areas of research. There would be a country—maybe continent—wide block party, for example, if research came out in support of "no differences" when comparing two groups:

- Group 1—Participants selected because they exercise, eat very healthful foods (e.g., occasionally drink a little too much, but it is carrot juice not alcohol), and do not smoke cigarettes.
- Group 2—Participants selected because they rest on their couches most days watching daytime TV or movies on their huge screens, nibbling french fries, or smoking cigarettes, while waiting for the delivery of their regular double pepperoni and lard breakfast pizza and bucket of double-coated chicken wings.

Just imagine the study—large sample size (great statistical power), no threats to validity, and dazzling measures of biological markers of cardio health and follow-up for 80 years to measure survival. What are our imaginary findings? Rejoice—"negative" (no difference) results. I can stop

bashing myself about living "Group 2." Forget the fantasy—there are real and realistic circumstances where we are eager to learn about no differences.

As an illustration, there is a serious food example where we would be extremely interested in a no-difference finding (i.e., no effect). This example focuses on genetically modified (engineered) versus non-engineered foods.

> As brief background, *food insecurity* refers to the availability of and access to food and is a worldwide problem.

There are many contributors to food insecurity, but two that are familiar are overpopulation of the planet (e.g., many more people to feed) and climate change (leading to a diminished area of land worldwide that can produce food). Worldwide, lack of available food is associated with many other problems (e.g., disease and death, political unrest, poverty).

> Also, there is something called the *food-insecurity-obesity paradox*. Countries with insufficient access to food (high food insecurity) have much higher rates of obesity in part because of the types of foods they consume (Franklin et al., 2012). (These same countries also have higher rates of malnutrition.)

Genetically modified food is one means of increasing and improving the supply because many characteristics of the food can be "controlled" (e.g., crops can grow faster, are more resistant to pests, are less subject to pests, are much more nutritious, and so on). The move to develop such foods holds promise to overcome worldwide starvation by providing increasing the supply of foods in heavy demand (e.g., rice, corn) and making these foods more nutritious. Many individuals who are not starving are still not receiving basic nutrients from the limited diets and have high morbidity and mortality rates as a result. Will genetically modified and nonmodified foods be exactly the same (no difference) on safety and side effect issues?

There is worldwide concern about the safety of the modified foods and one term to characterize the fear is "Frankenfoods" (modeled after the "Frankenstein") among some consumer groups (Busch & Howse, 2009; Kimenju & De Groote, 2008; Noussair, Robin, & Ruffieux, 2004). With this context in mind, we would all welcome replicated no-difference findings. Here is a definite case where "negative findings" would be very important. That is, we would want meticulously designed studies with humans that included measures in multiple domains of health and longevity to show "negative results." We would want these studies to be conducted by individuals not invested in the outcome because of funding they received from industry that sells modified foods. As well, we would want experimental nonhuman animal studies (e.g., to study biochemical processing of food, metabolism at the molecular level, psychological effects in learning, memory, perception). That evidence from human and nonhuman animal studies alone might not be persuasive because political and policy decisions are not always based on the best available science. And the best science often leaves many critical questions still unanswered. In the meantime, several countries are avoiding genetically modified foods until a clearer verdict is out, i.e., hoping "no difference" findings across a range of possible health outcomes. The decision to allow or not allow one's citizens access to genetically modified foods has dangers and risks on both sides. Not allowing such foods avoids unclear risks and dangers that many have voiced. Yet, in the process people are dying daily from food insecurity and poor nutrition and so there is more here than science, hypothesis testing, and what to do about the null hypothesis. Negative results could be extremely important in this area especially if we could trust them (well-designed studies) with the virtues of science (replication by independent groups of researchers).

15.3.3: Additional Examples of Negative Results Being Important

Consider the use of cell and smartphones as another example where negative effects would be important. There is some concern that use of cell phones may increase the risk of cancer, especially brain cancer, leukemia, and lymphoma. This is not a random worry. Cell phones transmit and receive radio waves, and these radio waves fall on the same part of the electromagnetic spectrum occupied by more powerful sources, such as microwave ovens and airport radar systems. Does the repeated use of cell phones influence (e.g., either increase risk or actually cause) cancer? Sadly, we know already that cell phone use is associated with increased rates of injury and death, but these stem from automobile accidents among those who drive and talk or text on their cell phone (Wilson & Stimpson, 2010). But what about cancer?

Here is another case where we would really love support for the null hypothesis, i.e., showing that cell phone use has no association with cancer and that individuals who do or do not use cell phones, controlling for other factors, show no difference in rates of cancer. That is, solid "negative results" would be quite important. The data are not so clear with one meta-analysis of several studies showing that use of a cell phone for at least 10 years is associated with an increase in brain tumors (Khurana, Teo, Kundi, Hardell, & Carlberg, 2009). The fact that the tumor is likely to occur on the side of the head as the one preferred for cell phone use adds just a little bit more plausibility that it is cell phone use. Yet another meta-analysis also spanning cell phone use for at least 10 years found no increased risk of brain tumors or cancer (Kan, Simonsen, Lyon, & Kestle, 2008). Science does not tell us how to act on available data and does not have "shoulds" associated with findings—even though implied recommendations (e.g., do not smoke cigarettes) might be obvious. In the case of cell phone use, one might to

be conservative and note that if several studies have found a cancer effect (from the meta-analytic review) that might be a guide rather than the negative results of other studies. That is, decisions in everyday life (but also in government and policy) are based in part on what would happen if one is wrong in the action one takes. People who are concerned about the cancer risk are encouraged to pursue one of two interventions—cell-ibacy (refraining from excessive use) or use of headsets, which allows speaking on the phone with the phone away from one's head.

The importance of no-difference findings can be seen in an abbreviated version of the tragic and ongoing story about vaccination as a cause of autism. Several years ago, a report suggested that the measles–mumps–rubella vaccine, commonly used with young children, may cause autism (Wakefield et al., 1998). This was not an empirical study but case studies of 12 children who had been functioning normally. Yet, they lost the acquired normal skills and showed behavioral symptoms of pervasive developmental disorder (autism now referred to as autism spectrum disorders in current diagnosis) as well as gastro-intestinal symptoms (from a viral infection). In the report, all of this was traced to the possible link between the vaccination and autism. To get to one of the punch lines, the article was identified as fraudulent years later and the connection between vaccinations and autism was bogus (Editors of *The Lancet*, 2010). The original publication set off a decade of public-health concern internationally, deep pathos and anxiety among parents of children with autism and among parents pondering routine vaccinations for their children, litigation against drug manufacturers, involvement of the U.S. Congress, and endless news media portrayals (Deer, 2011; Langan, 2011; Sugarman, 2007).

Long before the fraud was revealed and the original article retracted (in 2010), many studies of the putative link of vaccines and autism were completed. Also, in the United States, an Institute of Medicine panel evaluated the research available at that time date (Stratton, Gable, Shetty, & McCormick, 2001).[3] The conclusion: there was no link between vaccination and autism. That is, the research supports "negative results." Later reviews reached a similar conclusion (e.g., DeStefano & Thompson, 2004; Miller & Reynolds, 2009). Now hundreds of studies on the topic involving thousands of children suggest "no effect."

The issue and resolution convey the importance of clear scientific verdicts. The misery of parents who thought their own behavior (getting the child vaccinated) caused autism and the energy in fighting for their children (e.g., litigation, controversy) cannot begin to be represented here. The other part of the tragedy is that many parents (e.g., United States, United Kingdom) decided not to have their children vaccinated during the controversy to protect them from autism. Sadly, hundreds of children have died in each of the countries who otherwise would not have died had they had their routine vaccination.

The story is not over. Many parents are not persuaded that there is no vaccine–autism link and viewed key reports as riddled with conflict of interest and efforts to keep the vaccine policy in place at all costs. Moreover, many Web pages continue to fuel the flames by providing misinformation that there really is a connection between vaccination and autism (Calandrillo, 2004; Kata, 2010). Predictably, rates of measles, mumps, and whooping cough (pertussis) are up in the United States along with deaths of young children from these diseases (Gross, 2009).

From the standpoint of our discussion, the "negative results" were very important. It may be a matter of time before vaccinations are back to their prior levels and that is an important other topic about dissemination and diffusion of knowledge. Yet, for millions of parents and their children, the endlessly replicated "negative results" are very important.

The examples ought to make the case that no difference or "negative results" can be very important and indeed save lives (e.g., if we could get vaccination rates back up again). Perhaps less dramatic than the trauma-associated vaccine story is the importance of negative results or no-difference findings in other contexts. One of these pertains to cost. A controversial issue in the delivery of mental health services (e.g., psychotherapy) is who ought to provide services. On the one hand, trained mental health professionals with doctoral or master's degree are usually required (by the state law) to allow them to call themselves psychologists, family therapists, and so on and to administer treatment. There is now considerable research showing that one does not need an advanced degree to administer treatment effectively in the majority of cases in need and that there are no differences in outcome as a function of advanced degree or not (see Kazdin & Rabbitt, 2013). In fact, most developing countries do not have professionals to deliver mental or physical health care in proportion to what is needed. Novel models of delivery involving community individuals can administer treatment effectively, even for serious psychiatric disorders (e.g., Balaji et al., 2012; Patel et al., 2010). Here is a case where "no difference" in outcome is very important because the cost of delivering services is greatly reduced from what it would be by relying on highly trained professionals and more people can be treated. No-difference finding in outcomes? Very important.

15.3.4: Further Considerations Regarding Importance of Negative Results

No-difference findings may also be important because of the potential questions they raise. For example, programs that are extremely costly and mandated for rehabilitation for special populations (e.g., in prisons, special education classes) may be shown to have little or no impact in the outcomes they produce.

Such findings are critically important to know so that further resources are not wasted or used in ways that are likely to have minimal impact. A decade or so ago, millions of dollars were spent in the United States for programs that encouraged teens to take a virginity/abstinence pledge to postpone sexual activity. The goal was to decrease sexually transmitted diseases (e.g., HIV/AIDS) and unwanted pregnancies. In one evaluation mentioned previously, pledge and no-pledge matched teens were essentially no different on most measures of sexual activity 5 years later (Rosenbaum, 2009). (Actually, on some measures such as rate of unprotected sex, the pledge group was significantly worse.) The basic finding of no-difference was very important. Yes, we want our programs to be effective but as important is identifying those that are not so we do not waste resources and delay in developing and evaluating more effective interventions. Apart from health issues in this example, the more general point can be made. Negative findings can be very important when they have to ensure that resources are not deployed for ineffective programs.

In the context of domains other than intervention, negative results can be important as well. For example, researchers in one study devised a comprehensive assessment battery of motor skills (e.g., balance, walking) to identify among the elderly who was at risk for falling (Laessoe, Hoeck, Simonsen, Sinkjaer, & Voigt, 2007). The focus is important because falls among the elderly are a significant source of injury leading to disability and in some cases death. Also, many of the medications that are prescribed to the elderly (e.g., sedatives, antidepressants) further increase the risk of falling (Woolcott et al., 2009). If we could identify individuals at risk, there are training options that could help reduce that risk. The well-developed battery led to "no difference" (between those who fell and who did not). Negative results for sure are an important contribution to move research forward on other ways to develop other measures or indices that do make a difference. This is similar to the prior point about "no difference" findings for interventions. We would like differences, but we also want to know what does not work so that we do not continue to use that under the guise that it does.

In the general case, the value of negative results stems from both substantive and methodological considerations. The substantive considerations refer to the significance of the experimental or clinical questions that guide the investigation. As in some of the above examples (e.g., vaccinations and autism), the question may be one in which no differences between two groups, conditions, or interventions are actively sought. In many other cases where we "want" positive results, it is invaluable to know that something we are using and thought to be effective actually makes no difference. Progress depends on that.

> The methodological considerations refer to the care with which the study is planned, implemented, and evaluated. In particular, given documented weaknesses of research, the power of the study to demonstrate differences if they exist is pivotal.

Power analyses, of the type discussed earlier, can provide the investigator and reader of the report with a statement of the sensitivity of the test in light of actual effect sizes obtained. No-difference findings in a well-conceived and controlled study with adequate power (e.g., >.80) ought to be taken as seriously as any other finding.

15.3.5: Special Case of Searching for Negative Effects

> Randomized controlled clinical trials refer to studies in which an intervention is provided (e.g., medical or psychological condition) by assigning individuals to intervention (e.g., medicine, surgery, psychotherapy) as opposed to a control condition (e.g., no treatment, treatment as usual, placebo).

Trials for the treatment of various cancers, HIV, and other conditions in which life and death are involved often have careful monitoring of the data along the way before the study is completed. There are different reasons for monitoring the data in this way:

- Untoward side effects (e.g., including death) may emerge in one of the groups and we would want the trial stopped.
- The treatment condition is strongly emerging as more effective than some control condition and further assignment of cases to control condition ought to stop for ethical and clinical reasons.
- Our present discussion on negative (no difference) effects and their potential importance.

A clinical trial may be stopped if the intervention looks as if it is not worth pursuing further, i.e., is not working. This is referred to as futility analysis. (This is a formal term and not to be confused with my dissertation committee's comments that "further *analysis* of my dissertation was an exercise in *futility*.")

> *Futility analysis is designed to see if a treatment is unlikely to be better than a control condition* (DeMets, Furberg, & Friedman, 2006; Snapinn, Chen, Jiang, & Koutsoukos, 2006).

In relation to the present discussion, the goal is to identify whether there is likely to be a negative result or no difference. Among the reasons for doing the analyses is that conducting clinical trials often is very expensive. Also, in the case of medical procedures, the interventions often go through many phases to test for safety, then effectiveness, and so on. From start to finish, identifying an effective treatment can take many years. Also, there may be

many candidates for treatment (e.g., many different medications, many alternative and complementary treatments such as various herbs, vitamins, and minerals that look like they might hold promise) that are reasonable to test. It is not feasible to go through each one in a careful trial.

Two types of research usually are carried out to help sort through available treatments:

1. Animal model research to test (e.g., mice, rats) whether critical processes related to the disorder or disease might be affected by the medication.
2. Clinical trials with futility analyses are conducted to see if these interventions hold promise.

A trial is specified with careful methodological considerations (power, criteria for deciding when to look at the data and when to stop the trial). This is not a full-scale randomized controlled trial but something more abbreviated to see what might be promising.

As we see later in the chapter peeking at the data, before a study is completed to see how things are going often leads to an increase in "chance" findings as authors see "significance" and then stop or see nonsignificance and say we need more data. Methods of futility analyses go to greater lengths than the usual data peeking by specifying criteria for stopping and also trying to maintain the balance between identifying treatments likely to be futile or to be promising (e.g., Herson, Buyse, & Wittes, 2012; Jitlal, Khan, Lee, & Hackshaw, 2012). Futility analysis is mentioned in passing, although it is not used routinely in the evaluation of psychosocial interventions within clinical psychology and related mental health disciplines that evaluate psychological interventions. However, futility analysis is a case where negative effects are considered to be very important to identify and help research to move on to more promising interventions.

As in the case of methodology (and much of life again), invariably there are trade-offs. One might identify a treatment as futile but be mistaken (e.g., Type II error). And of course identify a treatment as promising when it proves futile (e.g., Type I error). Science and methodology are hardly error free—one is trying to make progress and place bets (based on theory, hypotheses, promising leads) on what interventions will solve a problem. Futility analysis is designed in part to do that more efficiently by determining if there are likely to be negative effects and to move on to more promising options.

15.3.6: Negative Effects in Perspective

The history of null hypothesis testing comes out of a tradition that emphasized falsifiability (Kragh, 2013; Wilkinson, 2013). That means that studies are designed in such a way as to provide evidence that the null hypothesis is false. One is required to "falsify" the null hypothesis rather than to "prove" the alternative hypothesis. From this view of science, one cannot prove the null hypothesis, and in this research tradition it still remains unwise to predict the null hypothesis (e.g., no difference)—we still operate largely on the notion of falsifiability, so we provide evidence to "refute" or be inconsistent with that hypothesis. But there is another reason too, namely, that showing no differences actually is very easy (but sometimes a lot of work as in my dissertation). Design a study with too little power (many such studies) throw in an unreliable measure or two, and do not monitor how the experimental manipulation was carried out and maybe toss in a couple of not too well-trained research assistants. Any one of these could wash out a true effect and show no differences. For these reasons—reliance on falsifiability and ambiguity of no-difference findings, negative results are more likely to be demeaned and considered not to "count" as a scientific contribution.

Common sense, logic, and even methodology might argue for a more favorable view of negative effects.

As for common sense, repeated demonstration (replicated effects) that no effect is obtained would argue for no real difference and support for the null hypothesis. One no-difference demonstration could be a fluke or a sloppy study; also a no-difference finding is "predicted" once in a while by "chance." Yet, another demonstration or two might be flukes or chance, but now we have a pile of no-difference findings. Common sense suggests that there is a hay stack and no needle, i.e., nothing to find.

As for logic, showing that repeated efforts to find a difference have failed does not necessarily mean an effect could never occur. That is, science does not make universal statements with words like "always" and "never." Logic teaches that universal statements are risky, unwise, and easily refuted with one instance. And despite repeated demonstration of no-effect, in principle one cannot say "vaccines never, ever could cause autism." This is not a statement about doubting the evidence we have but about science as a way of knowing, thinking, and talking.

As for methodology, at the beginning of the text I mentioned two criteria related to scientific explanations, namely, parsimony and plausible rival hypotheses. A repeated set of negative results (no differences) now raises these concepts. Is it more parsimonious and plausible to explain negative findings from a few or several different studies by a pile of individual explanations (e.g., this study had weak power, that study used sloppy measures, that study was with an odd population) or to explain all of the negative findings with one interpretation, "there is no effect." One has to look at each area research reporting negative effects, but the broad point can be made. The null hypothesis may be the more parsimonious explanation of a finding and plausible as well.

There is one more and arguably the most important methodological point to make. Null hypothesis testing and falsifiability is one approach to scientific research. It has been wonderful in the yield as well as controversial (all the business about statistical significance discussed previously and the point that groups are always different so statistical significance is a function of N). Yet, there are other approaches to do science that do not require rejection of the null hypothesis (e.g., Bayesian analyses, model building, qualitative designs). I mention this because we do not need to be rigid with statements like "negative effects" or support for the null hypothesis is not really interpretable. Not all scientists adhere to that and those who do are no less rigorous or informed than those who do not, i.e., probably "no difference" finding here too!

15.3.7: Further Considerations Regarding Negative Effects

In designing your own research, there are useful messages to draw from this discussion.

First, it is still the case that one ought to be very careful in the design and execution of the study. The threats to validity we discussed early in the text are not esoteric concepts—they have practical implications and influence each individual study.

For example, low power and excessive variability in the study (threats to data-evaluation validity) or diffusion of treatment (threat to internal validity) and others not mentioned here can operate to produce or make negative results more and sometimes very likely. An investigator cannot be held accountable for how the world really is (effect or no effect of my variable), but can be held accountable for the design and execution of a study and controlling or taking into account the many threats that undermine a clear demonstration.

Second, design a study so that not everything hinges on supporting a particular or single hypothesis. Expand the hypotheses to include a broader range of dependent variables that might make the findings especially interesting and make predictions about moderators.

A typical example is from my own work. We focus on improving the behavior of very aggressive children. We find changes in parents (e.g., reduction of stress, psychiatric symptoms) and the family (e.g., improved family relations) (Kazdin, 2010). Not earth-shattering I admit, but ancillary effects on important domains of family functioning did not alter our hypotheses (we made no prediction) but did influence the richness of the findings.

Third, go to greater lengths to ensure that the conditions you wish to compare are more starkly different experimentally.

Instead of comparing three groups (a little, a lot, and none) of some construct you care about (e.g., empathy, depression, love of methodology), start out your research by showing you can find a difference with the strongest comparison (a lot vs. none). Later you can become more nuanced as you master likely effect sizes, power, and other facets of running the study that will help as you wish to detect more subtle effects.

The conclusions from all of this: "negative effects" can be due to the status of things in the world, i.e., the variables really are not related. But it can also be the case that the variables really are related but you did not find that. Chance is one reason and we cannot do a lot about that—actually we can and that is taken up in the next section. Yet another reason is a weak experimental design. That part is in our court as investigators. A really well-designed study with a negative effect can be interesting and important, as some of the examples have conveyed.

15.4: Replication

15.4 Define the concept and role of replication in statistics

Evaluation of the results of an experiment, whether or not a significant difference is demonstrated, entails more than scrutiny of that one experiment alone. Can the finding be replicated, i.e., shown again in another empirical test? Replication or reproducibility of a finding is a pivotal topic because of its central role in science, the accumulation of knowledge, and evaluation of findings in relation to a particular study or demonstration. As a core topic, replication is not new. Yet, the importance of replication has received renewed attention in light of increased concern over nonreplicated findings, continued concern with the limits of null hypothesis testing, and scientific fraud, as we discuss here.

15.4.1: Defined

Replication refers to repetition of an experiment. This is a method of verifying scientific findings by showing or at least testing whether the original finding can be obtained again.

There are many different ways this can be done, but the basic concept is to repeat the study with the goal of evaluating whether the original finding is repeated. Replication of a study can be done by the original investigative team that did the study or by others not involved with that original study. Ideally, there will be multiple replications of a given finding and these will include replication by independent investigators. The credibility of a finding is enhanced when the replications go beyond the original investigator and her laboratory. With others show the finding, we are reassured that the finding did not depend on

procedures or facets of the original laboratory from which the finding emerged.

> Replication in principle is designed to make science self-corrective.

Over time, from repeated tests reliable findings will emerge and those that were due to chance, bias, some other fluke, or flat out fraud will drop out. That is, the very process (repeated studies, accumulation of information) will help correct the inaccurate or false results of any particular study or set of studies. As we will see, there are many competing forces against making the ideal real.

15.4.2: Types of Replication

Let us begin by clarifying what we mean by replication. There are many different types of replication, and there is no standard single definition or categorization that is universally adopted (see Schmidt, 2009). The different types vary in part as a function of how the replication study follows characteristics of the original investigation and the dimensions along which the replication effort may vary (e.g., types of subjects, tasks, means of operationalizing independent or dependent variables). Two broad types of replication are among the more commonly discussed and therefore useful to know. These are direct (or exact) replication and systematic or approximate replication and provide a useful way to convey critical points.

> *Direct replication refers to an attempt to repeat an experiment exactly as it was conducted originally.*

Ideally, the conditions and procedures across the replication and original experiment are identical or very close to that.

> *Systematic replication refers to repetition of the experiment by systematically allowing features to vary from those in the original study.*

The conditions and procedures of the replication are deliberately designed only to approximate those of the original experiment. It is useful to consider direct and systematic replication as points on a single continuum. That continuum might be labeled something like, degree to which the study resembles the original investigation. The top portion of Table 15.2 illustrates direct and systematic replication as opposite ends of a continuum.

A replication that is at the direct (left) side of the continuum would follow the original procedures as closely as possible. Procedures, measures, source of subjects (e.g., introductory psychology pool), and other features are very close to those used in the study one is trying to replicate. Obviously, direct replication is easiest to do for the researcher who conducted the original investigation, because he or she has complete access to all of the procedures, the population from which the original sample was drawn, and nuances of the laboratory procedures (e.g., tasks for experimenters and subjects, all instructions, data-handling procedures) that optimize similarity with the original study. Direct replication does not require the original investigation team, but is the extreme case where the replication is most likely to resemble the original study.

Table 15.2: Continua to Represent the Types of Replications: Useful Way to Consider, Distinguish, and Blend Types of Replication

Type of Replication	Description
Direct replication … … … … Systematic replication	Degree to which a study resembles the original investigation
Direct replication … … … … Systematic replication… … … … … Extension	Degree to which a study resembles the original investigation

NOTE: To facilitate understanding of types of replication, in principle one could place a given study at some point on one of the continua. The top portion focuses on clear replication attempts that try to reproduce the original finding but vary in how much the study departs from the original conditions of the study. The bottom portion expands the continuum. Extensions are hazy on efforts to replicate per se. Rather the goal is to see if the finding applies in quite different contexts. These extensions are sometimes called conceptual replications to note that they are not quite replications of the prior findings but replications of the concepts or principles that the original study supported.

Direct replication by someone other than the original investigator can be a little more difficult to conduct. Many of the procedures are not sufficiently described in written reports and articles of an investigation. Journals routinely limit the space available in which authors may present their studies. Hence, further materials about how the study was conducted usually must be obtained from the original investigator.

> Ideally, an individual interested in a close replication of the original experiment would obtain and use as many of the original experimental materials as possible.

Even so, many features of the study may be nuanced (e.g., who trained and supervised the research assistants, how much the investigator hovered over the execution). As discussed later, new efforts are underway to make direct replications easier to conduct by having the direct input and assistance of the original investigative team.

In principle, an exact replication is not possible, even by the original investigator, since repetition of the experiment involves new subjects tested at a different point in time and by different experimenters (researchers), all of which conceivably could lead to different results. Thus, all replications necessarily allow some factors to vary; the issue is the extent to which the replication study departs from the original investigation.

A replication toward the systematic (right) end of the continuum (top, Table 15.2) would vary the experiment deliberately along one or more characteristics. For example,

a systematic replication might assess whether the relation between the independent and dependent variable holds when subjects are recruited from the community (e.g., via MTurk or Qualtrics) rather than from a pool of introductory college students, when patient diagnoses differ, or when the therapists are inexperienced rather than experienced.

> A systematic replication tends to vary only one or a few of the dimensions along which the study might differ from the original experiment.

If the results of a replication differ from the original experiment, it is desirable to have a limited number of differences between these experiments so that the possible reason for the discrepancy of results can be more easily identified. If there are multiple differences between the original and replication experiments, discrepancies in results might be due to a host of factors not easily discerned without extensive further experimentation.

15.4.3: Expansion of Concepts and Terms

It is important to keep in mind direct and systematic replication as the key concepts and types. I mentioned that these can be clear at the margins (extremes) and as opposite sides of a continuum. Yet, it is important to expand this, in part because it may help you conceptualize your own studies and how to view other studies.

Systematic replication moves away from the original set of procedures intentionally by varying some dimensions or features (e.g., who serves as subjects, how the laboratory manipulation was implemented).

> At what point is that extension and departure from the original study not really a replication?

Great question. Consider a distinction between a *systematic replication* of research and *extension of a finding*. Replication is a term usually reserved for situations in which an effort is made to repeat the original study in a very close or fairly close approximation of the original conditions.

> When "replication" is involved the question is: Can that finding be demonstrated again?

The investigator intends to say at the end of the study that the findings support or do not support those of the original investigation.

An extension of an original finding begins with the view that some original basic finding is fine (and perhaps already replicated in direct replications). The extension is not aimed at testing the original finding per se. Rather, an extension focuses on external validity or the generality of the finding. When an extension or test of external validity is involved, the question is: Can the relationship among variables be extended to other samples, situations, settings, or circumstances?

What do you think?

The investigator intends to say at the end of the study that the findings from the original study also apply or extend to some other circumstance such as another clinical problem or population, cultural group, and so on. Failure to obtain results consistent with the original study is not necessarily an inconsistency or a failure to replicate. It suggests that the original finding may only hold under limited circumstances but the extension is not really a challenge that contradicts the original finding.

Consider now an expanded continuum (bottom Table 15.2) in which we have direct and systematic replication and now on the right side "extension." Systematic replication blurs with extension of the research as I discussed previously.

> Whether a study is more of a systematic replication with the main emphasis of reproducing an original finding *or* an extension, with the main emphasis on testing generality of a finding (external validity) is a function of how the investigator casts or frames the study.

The investigator may state the primary purpose of the study is a replication or extension. Also, a study can be both a systematic replication and extension by including the original conditions and novel conditions. For example, a study may have shown that depressed patients process information in a particular way. You come along and want to both replicate and extend. You have a hypothesis that processing information would apply to patients with obsessive-compulsive disorder. You include depressed individuals and test them the way that was done in the original study (replication part) but also include another patient population (the extension part) to test for generality.

The extension listed in the Table conveys that one is pursuing new ground and not just attempting to replicate a specific finding. Extension occasionally is described with another term, "conceptual replication."

> *Conceptual replication refers to a study that tries to demonstrate the primary relationship, concept, or principle of the original study* (Schmidt, 2009).

A conceptual replication makes no pretense of being a direct replication, so there is no possible confusion there. Yet, conceptual replication can encompass any test that abstracts the principle or guiding concept of a study (e.g., how calorie-restricted diet affects another species; how social norming or priming of behavior applies in nonexperimental naturalistic situations). I will give examples later and refer back to direct, systematic, and extensions (conceptual replications). The continuum is a useful aid to summarizing key concepts and their relation.

15.5: Importance of Replication

15.5 Explain why replication is important in scientific research using the five key reasons

The importance of replication in scientific research cannot be overemphasized. In many ways, replication is the backbone of all of science. Table 15.3 summarizes key reasons why replication is so central and these reasons are elaborated further here.

Table 15.3: Key Reasons Why Replication Is so Important

Key Reasons
1. Reasons 1 and 2 for the Importance of Replication: To help eliminate or reduce the probability that the initial finding occurred by "chance"
2. Reasons 1 and 2 for the Importance of Replication: To make less plausible that some unclear influence may have operated in the original study or that some confounding variable (e.g., characteristics of the subjects) contributed to the difference
3. Reasons 3, 4, and 5 for the Importance of Replication: To make less plausible that special ways in which the investigator handled the data or measures or made decisions directly influenced the finding (e.g., whether and how to delete outliers, which measures to include, which analyses were and were not reported)
4. Reasons 3, 4, and 5 for the Importance of Replication: To ensure the original findings were not based on fraud by data fabrication
5. Reasons 3, 4, and 5 for the Importance of Replication: To sustain credibility of science and scientific findings to the funders and consumers of research, namely, the public. Decisions are made on scientific findings, and we need to be sure those findings as are as solid and replicable as possible.

Each of the above reasons is discussed in detail in the following sections.

15.5.1: Reasons 1 and 2 for the Importance of Replication

Here are two key reasons why replication is so important.

First, replication is essential because of the prospect that any single finding is a result of "chance," a feature inherent in the nature of null hypothesis testing and statistical evaluation. As investigators, we invariably assume that when we find statistical significance, this means the finding (relationships among variables) is real and if the p level is really low ($p < .0001$), then this is very real and clearly beyond chance. Yet, any finding might be due to chance and the special sampling of subjects under the circumstances and time the study was completed. Chance becomes very much less plausible upon replication of a finding in a new study and of course implausible if there are multiple replications.

Second, replication is important because many influences might operate in an experiment that could lead to a particular pattern of results. The findings may reflect a "real" ("non-chance") effect but some other influence may operate to account for that effect (threat to construct validity).

The influence might be subtle (subject or experimenter effects) or not so subtle if a confounding variable can be identified. For example, let us say we show that a group of nonsuicidal self-injury participants vary in their pain threshold from controls without a history of self-injury. Yet, our finding is not replicated. It could be that nonsuicidial self-injury in our study was confounded with another variable (e.g., clinical depression, parental history of suicide) and that was the "real" reason or contributed to our finding. Our finding was not chance but a real finding. The replication attempt included subjects without the confounding characteristic. A successful replication of the results does not necessarily rule out such influences but increases the likelihood that the independent variable or experimental condition is the common feature across the multiple studies that explains the relation.

15.5.2: Reasons 3, 4, and 5 for the Importance of Replication

Additional reasons regarding the importance of replication are outlined below.

Third, and related to influence that replication can combat investigator biases related to how the data are analyzed and presented and what decisions are made by the investigator in preparing the report of the study. I mentioned in the discussion of negative effects that there are many decision points in data analyses, selecting among the measures to report, selecting among the many data analyses and subanalyses, excluding some data points or "outliers" depending on the impact of such exclusions on the data analysis, and other choices that the investigator makes in selecting and presenting the data (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). For example, there may be scores of exploratory analyses, but those that obtain significance are used to confirm what the investigator may want to show. As one set of authors aptly put this, in studies we "torture the data until they confess" (see Wagenmakers et al., 2012, p. 633, for more on this quote). Consider one practice investigators occasionally engage in that could lead to a spurious finding.

While the study is still in progress and participants are being run, the investigator may take early peeks at the data. This means that the full set of participants who are to be included have not yet completed the study. Yet, the investigator conducts some preliminary data analyses to see how things are going, i.e., whether an effect is "coming out." Decisions then are made (do we really need more subjects, we already have an effect). As it turns out, this early data peeking and decision-making introduce a greatly increased risk of finding and reporting a chance effect (Francis, 2012).

We need replications to see that the finding was not due to selective reporting or other special, biased handling of the data. Here too we must be sure that the replicated study is not tinkered with in the same way to produce a selective view of what was obtained.

> Can we categorically say that early peeking is statistical voyeurism and shameful?

No, not at all, as discussed previously in the context of futility analyses. In many randomized controlled trials (e.g., treatment of cancer, HIV/AIDs), early peeking is planned, and once a verdict seems clear, the trial is stopped for ethical and clinical reasons (e.g., there is no longer a justification for placing individuals in an ineffective treatment or placebo-control condition or some predetermined criterion was met to suggest that the intervention is not going to work). Yet, in the general case in psychology experiments, the full study ought to be run because early peeking at the data is used selectively to capitalize on finding an early chance effect.

Fourth, replication is needed to ensure that the findings are not spurious because of fraud! Occasionally results of a study are reported where the data are faked. I have mentioned one instance in relation to vaccination and autism. Fraud can include a variety of acts such as saying that subjects were run when they were not, making up or altering the data, or reporting something that is simply not so. Fraud in scientific research generally is regarded as rare, but that is hugely important whenever it occurs. Apart from the specific finding, there may be enormous consequences for the public. That is evident on the faked vaccination-autism issue where public health issues emerged and are still evident with more children ill and dying because their parents understandably avoid vaccination of their children. The public trust was lost and in this that translates to loss of children's lives—absolutely tragic.

Fraud undermines science and public trust in science even when life and death are not involved. It could lead to cynicism about the entire enterprise and well beyond the particular instance of fraud. For example, the work of a prominent social psychologist from the Netherlands (Diederik Staple) was accused and found guilty of faking over 50 publications. He was suspended from his university in 2011; a final report investigating his behavior was completed in 2012. From accusation to final adjudication, the case was made very public in many major print and online outlets, apart from professional journals. In the public media, the accusation was made that perhaps all social psychological research and more generally all psychological research should be mistrusted. And in other places "Dutch universities" were indicted too as in question. There was no reason to extend the fraud suspicion to social psychology, then to psychology, and to all Dutch universities. Yet, reason is not the driving force. Again referring back to the vaccination-autism case, public trust is much more easily lost than won back and the autism mistrust is alive and well. Replication can help to establish reliable and veridical findings.

Finally, apart from fraud, replication can increase the credibility of science and scientific findings to the funders and consumers of research, namely, the public. Decisions are made all of the time that rely on the best available science (e.g., foods made available in the schools, safety standards for pollution, how to alleviate climate change, how to handle trauma from natural disasters, how to educate and care for children). It is fine for the news media to publish a catchy finding here and there, but we have to be cautious. Because of the reasons noted previously, a given finding may not hold and there are dramatic examples, as illustrated below. We want to be assured that information circulated to the public is based on replicable findings.

There are various reporting organizations designed to provide unbiased summaries of findings in a given area (e.g., Cochrane Reviews [www.cochrane.org/cochrane-reviews], Institute of Medicine [www.iom.edu/]) so mechanisms are in place to consolidate findings from many studies. There may be limits because reports have to draw primarily on published articles, and the publication bias may contribute to what we believe we know. But more in the day-to-day trenches, we need to replicate findings in our field, be encouraged to do so, and have publication outlets when we replicate and fail to do so. More inconsistencies might well appear in the journal literature this way, but fewer spurious findings might be published.

15.5.3: Instructive but Brief Replication Examples

A few examples in different contexts can convey key points about replication, but also other issues we have discussed such as "negative effects" and their importance. Also, with just a few examples we can squeeze in illustrations of direct, systematic, and conceptual replications.

> **Supportive Treatment for Cancer Patients:** Several years ago, an extraordinary finding was reported by a well-established set of investigators.

I mention "well established" not to argue from their authority, but to convey the group was well experienced in conducting studies, evaluating data, and other features that comprise the topics of this text (Spiegel, Bloom, Kraemer, & Gottheil, 1989). Patients (N = 86) with metastatic breast cancer were randomly assigned to supportive treatment versus routine care. All patients received routine care normally provided to oncology

patients. The supportive treatment condition was a form of group therapy with sessions that focused on building social bonds with others, expressing emotions, obtaining support from family and friends, and learning self-hypnosis to control anxiety and pain. The supportive treatment improved quality of life of the patients, but it is the quantity of life that was the stunning finding. At follow-up 10 years after the treatment began, all but three patients had died. By this time, the duration or time of survival of individuals from each group could be evaluated. The groups showed stark differences in their survival. Those who received the supportive treatment survived a mean of 36.6 months; those who received routine care without the supportive treatment survived a mean of 18.9 months. Supportive group therapy almost doubled how long people lived over and above standard care. This is a proof of concept study (can a psychosocial intervention lengthen lives among cancer patients) but also a clinical study that goes beyond concept and lengthens human life. What a stunning outcome!

Now to the punch line. Many replication attempts have been conducted, and they have been consistent in showing there is no survival benefit. The original team completed one of the replication attempts, which means we are as close to a *direct replication* as we can hope for (Spiegel et al., 2007). The results showed no statistically significant difference in survival for the supportive treatment and routine care groups (median = 30.7 months and 33.3 months, respectively). (Do not be distracted by the fact that the control group survived "longer." No difference statistically between these means the "difference" is well within normally expected fluctuations and that the groups come from the "same" population—whether treated or not.) As I noted, there now have been many replication attempts, some controversy about issues in the original study and replication efforts (e.g., see Coyne, Thombs, Stefanek, & Palmer, 2009), but the evidence suggests consistently that supportive group treatment is not likely to have impact on the length of survival. The original goal was to improve the quality of patient life and that finding stands but living longer is likely to require another intervention. The many replications (direct and systematic) obviously were critically important. We do not want a negative finding here and are eager to have all possible options for patients (for ourselves) to improve quantity and quality of life. Better to know what not to do and pursue other options. Negative findings and replication were hugely important here.

> **Parent Management Training as an Evidence-Based Treatment**: PMT refers to procedures in which parents are trained to alter their child's behavior in the home (Kazdin, 2005).

The parents meet with a therapist or trainer who teaches them to use specific procedures to alter interactions with their child, to promote prosocial behavior, and to decrease deviant behavior. Parents are trained in a variety of concrete procedures (e.g., how to use antecedents, how to shape behavior, special way to praise, and many other techniques that draw from an areas of research referred to as applied behavior analysis).

> The intervention has been quite effective with children who engage in oppositional, aggressive, and antisocial behavior (e.g., fighting, lying, stealing, firesetting, running away, confronting others).

In terms of psychiatric diagnosis the more extreme conduct disorder includes the severe behaviors; oppositional defiant disorder is less extreme but can still be hugely challenging. Both disorders have unfavorable long-term outcomes with high rates of psychiatric disorders and physical health problems too into adulthood.

PMT has been extensively studied and with slight variations in age of children (e.g., 2 to 18), with varying degrees of severity of dysfunction (inpatient children, outpatient cases), and more. A now dated review identified 82 studies of variations of PMT that included more than 5,000 children in support of the treatment (Brestan & Eyberg, 1998). Since then, scores of additional randomized, controlled studies of PMT have been completed with youths varying in age and degree of severity of dysfunction (e.g., oppositional, aggressive, and delinquent behavior) (Kazdin, 2015; Michelson, Davenport, Dretzke, Barlow, & Day, 2013; Weisz & Kazdin, 2010). Treatment effects have been evident in marked improvements in child behavior on a wide range of measures, including:

- Parent and teacher reports of deviant behavior
- Direct observation of behavior at home and at school
- Institutional (e.g., school, police) records

The effects of treatment also have been shown to bring problematic behaviors of treated children within normative levels of their peers who are functioning adequately in the community.

One might ask, "Do we need that many replications beyond the 82 reported in the late 1990s?" The replications have made many extensions as for example, to now include both treatment and prevention, and children with different types of problems (e.g., autism spectrum disorder, hyperactivity, anxiety). And beyond the present scope, many of the studies focus on moderators and mechanisms of change. For the present discussion, it is important to that PMT is a well-established evidence-based treatment. In this case, well established derived from direct and systematic replications and now many extensions (conceptual replications) well beyond the original foci (see Kazdin, 2015). Here the replication of positive findings is also hugely important.

15.5.4: One Additional Replication Example

Consistent Patterns of Behavior: Fixed-Interval Responding: B. F. Skinner (1904–1990) is one of psychology's luminaries in the area of learning.

Skinner's work on principles of operant conditioning began with meticulous animal work (rats, pigeons) and elaborated all sorts of facets related to reward, punishment, and more.[4] Much of this work has been developed and broadly extended to interventions that have been applied to:

- Virtually every age group (e.g., toddlers to the elderly)
- Patient populations and problems (psychoses, eating disorders, addictions, tics)
- Settings (schools, business)
- Contexts (e.g., military training, professional and amateur athletics) (see Kazdin, 2013a)

In an animal laboratory context, rats and pigeons can press a lever that leads to reinforcement (delivery of a food pellet).

The rule about how many lever presses and when food pellets are delivered is called the *schedule of reinforcement*. There are many schedules, and they lead to different patterns of behavior. We are considering just one referred to as fixed-interval responding.

When the schedule is predictable and the time interval is constant or fixed (e.g., every 10 minutes), the press of a lever after the interval elapses leads to a food pellet. Lever presses before the interval elapses do not speed up the delivery of food—the interval has to elapse. There is a very predictable pattern of responding. Right after reinforcement (pellet), there is not much activity (pressing the lever) but as the end of the 10 minutes approaches performance picks up and reaches a very high rate close to 10 minutes at which point the pellet is delivered and performance again drops off.

At this point, it would be reasonable for you to think, "who cares?"

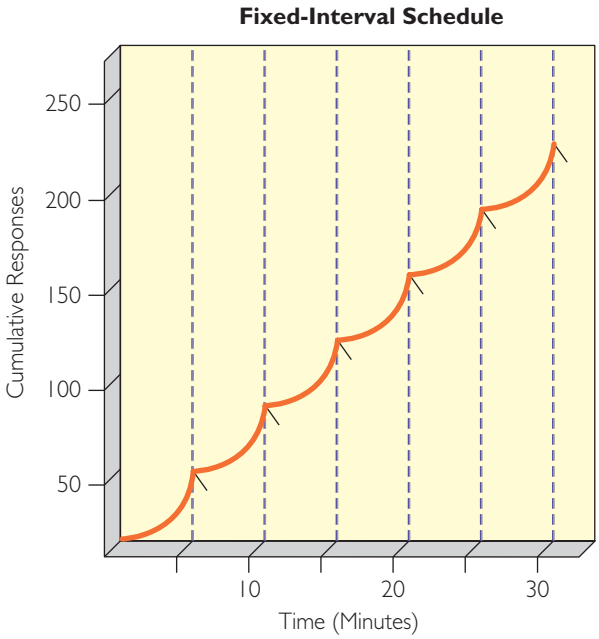
Ok before getting to the replication point, consider your own behavior.

Rather than lever pressing, which you do not seem to be doing too much these days, consider your reading behavior before an exam. Let us count number of pages read or hours of studying. There is no "food pellet" but there is a fixed interval. The day of the midterm and the day of the final exam are the end of the intervals. It is the beginning of the term, and you are not reading much for the midterm exam—the end of the interval is far away. As an exam approaches (i.e., the event on a fixed interval), the number of pages you read or the number of hours you study is likely to really increase and peak (be the fastest, most intense) right before the exam (interval). After the exam (e.g., next day, and days after that), your reading for that course is likely to stop and only slowly begin again, but to get faster as you get near the end of the next interval (final exam). The pattern evident here is pretty much what is evident in many different contexts.

Figure 15.2 shows a graphical display of what responding looks like when there are fixed intervals. Again the pattern—faster and faster responding as the interval nears. Little or no responding right after the interval and then start up as the interval gets closer and then faster and faster again. Who shows this pattern? Laboratory studies with rats, pigeons, and monkeys show the effect in the context of behaviors such as lever pressing. Humans show the behavioral pattern in that context too with lever pulling or pressing for some reinforcer. Yet more interestingly well outside the laboratory, the characteristic pattern is evident.

Figure 15.2: A Hypothetical Cumulative Graph Showing Fixed Interval Responding

A cumulative graph plots performance in an additive fashion over time.



NOTE: The score or data point (e.g., number of lever presses) at one point in time is added to the value of the next score or data point plotted on previous occasions. The score may take on any value of the dependent measure. Yet the value of the score that is plotted is *the accumulated total* for that day plus the sum of all previous days. This means when looking at the graph a steeper slope means more responding (faster) at those times and a flat (horizontal) slope means there is little or no responding.

For example, the pattern is reflected in the rate at which the U.S. Congress passes bills. At the end of a Congressional session (fixed interval), bill passing (rate or number of bills) shows the characteristic schedule effect (faster bill passing)

and this tapers off to little or no responding immediately after the interval when the new session begins (Critchfield, Haley, Sabo, Colbert, & Macropoulis, 2003). The message is not that some of the behavior of the U.S. Congress is pretty much like the behavior of rats and pigeons. Rather, the point is about the replicability of fixed-interval responding. These are direct, systematic replications as well as extensions (conceptual replications). The effect has been found in so many different contexts, with so many different subjects, and so on. Replication has gone well beyond showing the effect could be repeated.

> **General Comments**: In general we want to know if these scientific findings were replicated. Despite the universal importance of the question, the fact is that most scientific findings are not tested to see if they are replicable or at least we do not know that they are tested because they do not make it into the scientific literature (Pashler & Harris, 2012).
>
> There are notable exceptions where replications follow a study and sometimes relatively as quickly. Three conditions in particular are likely to generate replication studies. These include when a topic:
>
> 1. Is related to a great health concern (e.g., support groups for cancer, vaccines, and autism)
> 2. Has enormous scientific or theoretical consequences (e.g., first time an animal was cloned)
> 3. Violates a prevailing theory or consensus about what we know (e.g., whether neutrinos surpass the speed of light)
>
> These three conditions are not independent. They convey a bias so to speak of what type of study is likely to receive attention for replication.

Overall the rate of replication is recognized to be low. For example, one review examined the publication history of 100 psychology journals and searched for the word "replication" in the text (Makel, Plucker, & Hegarty, 2012). Of all psychology publications in this sample used the term, 1.6% included the term "replication." A more in-depth analysis of 500 randomly selected articles showed that 68% of the articles that used the term "replication" actually were replications. This would reduce the replication rate to 1.07% of studies.

There may be more replication studies than we know. What counts as a replication (e.g., on the direct—systematic—extension continuum) is not always clear. Also, any individual study might include a secondary hypothesis or finding that is presented in an effort to replicate. Negative results tend not to be published very much, so failures to replicate are not easily identified. The publication bias operating for the original study that favors a positive finding also is operating for the replication efforts. So the publication bias feeds the dissemination of chance findings and then their support by yet another finding that may be chance. So when we see two or five studies that replicated an original finding, any sigh of relief or assurance may be misplaced. It could be that scores of negative results were not written up, those that were may not have been submitted for publication, and those that were submitted may not have been accepted.

The selective reporting of data and data analyses raises a broader issue. Many experiments are completed and yield findings that are not statistically significant. The results of such experiments usually are not reported but merely allocated to a file drawer. The *file drawer problem*, as this is sometimes called, refers to the prospect that the published studies represent a biased sample of all studies that have been completed for a given hypothesis (see Rosenthal, 1979). Those that are published may be the ones that obtained statistical significance, i.e., the 5% at the $p < .05$ level. There may be many more studies, the other 95%, that did not attain significance. Computations can be completed to estimate how many unpublished findings without significant effects (i.e., the "file drawer") would be required for the finding in the literature to be challenged (see Ferguson & Heene, 2012; Pautasso, 2010; Rosenthal, 1991).

> This is also referred to as a *fail-safe number or the minimum number of unpublished studies that show no effect that would be needed to overturn the conclusion from the available published studies*.

Areas of research on the effectiveness of interventions in mental and physical health care (e.g., psychological treatment for PTSD, exercise impact on the elderly) often evaluate whether the impact of treatment from a summary from available published research is likely to be influenced by studies languishing in the file drawer (Peterson, Rhea, Sen, & Gordon, 2010; Powers, Halpern, Ferenschak, Gillihan, & Foa, 2010).

> **The goal is to answer:** If these (published) studies are likely to be artifacts (chance findings), how many nonpublished, file-drawer studies with negative effects would it take to challenge the conclusion that there is an effect?

For example, thousands of studies attest to the effects of psychotherapy for children, adolescents, and adults. It would take several thousands more studies with no effects in some file drawer to contest this basic finding. Consequently, it is not very plausible that the effects in the published studies are due to chance or biased reporting. As a general rule of course, as more studies support a particular finding, the less likely that any unpublished findings would negate the overall relationship that has been found. Analysis of the file-draw problem and how, whether, and the extent the analyses suitably address the publication bias, and the exact methods for computing the file-drawer number are a matter of controversy (see Francis, 2012; Simonsohn, Nelson, & Simmons, 2013). What is important to know here is that there have been efforts to take into account the scope of unpublished work that would be needed to challenge a finding.

15.5.5: Renewed Attention to Replication

Challenges for Identifying Reliable and Reproducible Findings: Replication occupies a special place in science and is important for all the reasons discussed previously. We want findings replicated and to know when they are not.

Many challenges present obstacles from simply accumulating findings to clarity of what we know through replication.

First, the publication bias is perhaps the greatest challenge for identifying reliable and reproducible findings.

As I mentioned in the discussion of negative results, journal publication greatly favors articles with "positive effects," i.e., studies where the null hypothesis (no effect) was rejected based on statistical tests. This means that many articles that are published might well be those chance effects.

What is the scope of the problem of the publication bias?

There is no firm way of knowing, but different authors have reached the dramatic conclusion, occasionally supplemented with mathematical proofs and simulations that many and even most publish research findings are not correct, i.e., are false (see Francis, 2012; Ioannidis, 2005; Moonesinghe, Khoury, & Janssens, 2007). Some estimates of how many efforts there are to replicate findings have encompassed very broad ranges (e.g., ~4% to 50% of studies are replications) (see Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002; Wagenmakers et al., 2012). The findings cover different sciences and areas of work in a given science. More importantly, there is no standard definition of what would and would not count as a replication. This does not alter the main point—publication bias interferes with our identifying what findings are and are not replicable. This concern spans the sciences (e.g., genetics, molecular biology, epidemiology and public health, psychology, and others).

Second and related, I mentioned before that "chance" will lead to positive findings once in a while (referred to as Type I error).

Yet, Type I error is likely to greatly underestimate "chance" findings if investigators selectively report analyses they completed, peek early at the data and make decisions about whether to stop the study then, and so on. We have focused on positive findings that are not "real" because they occurred by chance, decision making during the analysis of the study, and bias in the publication process. There is the other side and arguably even a greater problem, namely, negative (no difference) effects are likely to occur by chance too (Fiedler, Kutzner, & Krueger, 2012). That is there is a real effect, but we did not detect that in our study (referred to as Type II error). This is a great problem because, as discussed previously, most studies do not have sufficient statistical power to detect the difference that might be evident in the phenomenon (between groups). This means that we may be routinely discarding hypotheses and findings because we concluded that there was nothing there. We can do some estimation of the likelihood of Type I and Type II errors, but not as precisely as we think because of many indeterminancies of what gets studied, filed, written up, and communicated to others (e.g., John et al., 2012; Simmons et al., 2011).

Third, a great challenge is the mixed message given (in textbooks, classes on methodology) about the importance of replication.

On the one hand, we say, as I have done, that replication is the backbone of science and that we prize evidence that findings are replicable and reproducible. On the other hand, there are multiple pressures to researchers young and old to do something original—maybe come up with a new little theory along the way and test it. At some universities, there is a pressure to publish. In psychology, that means publishing in peer-reviewed journals, i.e., where publication bias rules. Publishing in such journals is not so easy because the journals that range from decent to very prestigious reject the vast majority of manuscripts (e.g., often ~80%) that are submitted. (That is why I submit papers to journals that range from no standards to desperate for articles.) At other universities, the pressure is beyond publishing articles; it is about impact, i.e., publishing novel ideas and findings that will shape the rest of the field in light of their innovativeness and importance. At no university (to my knowledge) is there the slightest interest, yet pressure, to replicate. What would you tell a graduate student or budding faculty member? "Do something that has already been done, that no one will regard as new, and that will be close to impossible to publish." Obtaining grants, job promotion, invitations to conferences, and so on are more of the same—replication really is not encouraged and actually has been discouraged. The mixed message (replication is so important, but it will not help you personally in any way or be of much interest) is like telling someone never to steal but giving a big wink and a smile while saying that.

These challenges do not change what we need to accomplish, namely, to identify replicable findings and to be sure the foundation we build is solid. Expected challenges of replication (chance of Type I and Type II errors) are exacerbated by publication bias. What can be done? I thought you would never ask.

Activities and Remedies to Support and Increase Replication: There are several suggestions to address the problems of publication bias and to improve the likelihood of our findings are replicable.

Many of these are listed in Table 15.4. It is important to know them and to incorporate those that apply to individual studies. No one suggestion solves the issues of guaranteeing

Table 15.4: Recommended Procedures to Improve Research That Have Implications for Interpreting Negative Effects and for Replicating Studies

Procedure	Implications
Include an effect size measure as part of the data analyses	This is wise for several reasons but has been suggested in this context to rely less on the binary significant/not significant decisions that null hypothesis testing has fostered
Conduct meta-analyses	Pool studies to evaluate effect sizes and as part of that computed confidence intervals of the likely range for the real effect
Specify primary measures, methods of analyzing data, and decisions in advance of executing the study	Post hoc searching for significance, identifying which measures or scales are "primary," ought to be clear at the outset so that later analyses and decision making do not pluck significant and more likely chance findings selectively
Improve/strengthen reporting standards for authors	To encourage reporting of all measures that were in the study, all analyses, and all other decisions that might influence what was reported
Keep confirmatory analyses (tests of the original hypotheses) distinct from exploratory analyses (fishing to learn more)	When exploring the data is used to confirm the hypotheses, this is a key place where decision making leads to biased findings. Specify measures and tests in advance. Additional findings and tests to explore the data and to generate hypotheses for future studies should be distinguished so that exploration does not mascaraed as confirmation
Move away from null hypothesis testing	Increase the use of other models (e.g., Bayesian analyses, qualitative methods)
Conduct and report multiple studies testing the same or related hypotheses	It is less likely that biased decision making across all the different studies will lead to (replicate) the same finding if that is "chance." Also, multiple studies means at least replication in some key ways (e.g., systematic replication). This is more feasible in laboratory (e.g., with college students or mice) rather than clinical studies) where access to patient populations is more restricted
Compute file-drawer	Calculate how many null (negative effects) it would take to jeopardize the positive results that have been obtained
Alter the policies of publication outlets to better accommodate replications and/or "negative results"	Have existing journals allocate space specifically to replication studies and/or "negative results." Develop new publications that will allow such studies in much greater numbers
Alter the incentive structure for scientists to promote replication studies	Having one's work cited (by other researchers) is related to job promotion and other career issues. Replication work, if done, rarely gets cited. Connect replications to original studies (e.g., electronically) so studies are "co-cited." Thus, an author completing replication research will not have a disincentive because the work will be neglected
Draw on students in training for replication research	Many students are available, developing their research skills, and looking for projects. This could be an untapped resource for doing replication studies integrated into course work or outside of class projects
Accept studies for publication based on the quality of the methods	Decisions can be made on methodology rather than outcome
Routinely make data and methods available	As part of any study, there should be open access to information (methods, measures, data) would permit further or reanalyzes of the studies and also use of procedures for replication
Educate reviewers	The peer-review system might be better informed to tolerate mixed results (not all positive findings) when making decisions to accept or reject a paper
Devise a model or mechanisms that formally advances replication studies and actually checks on the reproducibility of findings	Something more formal is needed to provide a mechanism of replicating studies. The Reproducibility Project is discussed in greater detail in the text as an example

that we have replicable findings. Each is not discussed here now that you understand the challenges noted in this chapter and some of the suggestions already listed in the table. Two suggestions warrant a bit more discussion because they are changing research and how that research is conducted and reported.

No one suggestion is designed to redress the issues related to publication bias, "negative results," and the reproducibility of findings. Each solution has its individual strengths and limitations, and their discussion is beyond the scope of this chapter (see Asendorpf et al., 2013; Fiedler et al., 2012; Frank & Saxe, 2012; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2012; Schmidt, 2009; Schimmack, 2012; Simmons et al., 2011; Wagenmakers et al., 2012).

15.5.6: Additional Information Regarding Renewed Attention to Replication

A Novel Journal Publication: A recent journal in psychology has emerged that is unique in several ways related to interpretation of data, negative effects, and replication.

The journal, *Archives of Scientific Psychology*, publishes articles in any area of psychology (Cooper & VandenBos, 2013; www.apa.org/pubs/journals/arc/index.aspx). The articles are free and open to the public. The novel feature of the journal is that the authors will be required to complete a questionnaire based on already available journal article reporting standards (see JARS & MARS,

please see For Further Reading). The questionnaire will require a very explicit description of the rationale, methods, results, and interpretation. Also, authors will be required to make the data for the study available to others. Other scientists but also the public at large will have access to the articles, so authors are required to prepare some of the materials (Abstract, Methods) in both technical and more readable, lay versions. Articles may be published with comments from reviewers as well as replies or further comments by the authors. In addition, the Web site allows public posting of comments and discussion.

The journal addresses several issues that have been impediments to science, not just psychological science:

1. The detailed report (questionnaire) will provide much more information about a study and how it was conducted than publications normally allow. As I noted, replication can be difficult because so many procedural decisions are made that cannot be described in journal articles. The questionnaire will make replication more feasible in principle and practice.
2. Decisions also can introduce biases in studies that we have discussed in detail (e.g., in selecting or using some measures, some subjects, and some data analyses among of all those completed). Making the decision process and the rationale for decisions explicit can help combat these influences.
3. Information from evaluation of the study (by reviewers) and discussion from anyone interested who cares to be involved allow for multiple perspectives on all facets of the study. All can comment on credibility of what was done and the findings, make suggestions for improvement, and in the process guide further research.
4. The data for the study will be available. An incentive for authors to make data available is that any use of the data by others will require the new investigators to include the original investigators as authors on any paper. Thus, reticence of investigators to turn over their hard-earned data may change with the incentive of sharing in any publications that have used that data.

Overall, and perhaps most importantly, the journal fosters transparency at multiple levels. What the author did and the thinking behind that are clarified. Comments of those who reviewed the paper are available and so immediately we have potentially different perspectives on what was done (methods, procedures) and what was found. Others including the public and their views can be provided as well. This very open strategy may not eliminate bias, but certainly places findings in a context in which bias is reduced and discussed.

15.5.7: The Reproducibility Project

I have discussed replication as the backbone of science. Yet, the ideal is impeded by publication bias, chance positive findings in high rates, chance negative findings that are neglected, and no feasible way to evaluate the extent and scope of nonreplications. Transparency at all levels is perhaps a precondition for all of science, so there can be replication. This new journal has developed to redress many of the problems we have discussed in this chapter.

> ***Reproducibility Project*:** If we want to know about whether findings are replicable, the most direct way is to do the replication studies.

Researchers (e.g., psychologist and Nobel laureate, Daniel Kahneman) have suggested a ring or network researchers who engage in replications (Yong, 2012b). The Reproducibility Project (and abbreviated here as R Project) goes further by developing precise procedures regarding how this can be accomplishedwithin psychological research (Open Science Collaboration, 2012). (For this project, "reproducibility" is used interchangeably with "replicability.") The project has solicited researchers who participate on a team that will design, develop, and run a study to replicate prior work.

Multiple studies were selected from prominent journals, and from these a smaller set of studies were selected for replication. For any given article, the focus is on a key finding to replicate and ensure that the replication will have acceptable statistical power (>.80) to identify an effect if there is one. An effort was made to provide a direct replication in which the materials from the original experiment are used, if available. Also, the authors of the original study are involved to provide feedback on the design of the replication study and identify in advance of running the study what factors might interfere with the replication attempt. Changes the original authors recommended can be integrated into the study design and the replication study can begin.

The R Project is underway. It will be interesting of course to learn about the replication of individual projects. Yet the significance of the R Project stems from several broader considerations:

1. The project includes several senior researchers and to give their stamp to the importance of replication research. We have had years of demeaning such research and the mixed message of why do it at all has been more one-sided (do not do it) than mixed.
2. The project makes replication an empirical issue, i.e., data will be collected on many different studies. There have been scores of simulations some of which I have stating how many or most findings are likely to be "chance" in light of the bias to publish mostly "positive results." The R Project can provide real data and what findings and how often findings are replicated.

3. The data will not only include the results of individual replication studies but also a look at what might be involved in successful replication studies. That is, are there factors (e.g., characteristics of original studies, topics) that influence the replicability of a study? This is a higher level of abstraction that looks at replicability more generally.
4. By working out the details of how to proceed, the R Project has provided a model for doing replications and in a transparent and defensible way. The model consists of constructing teams, developing a design with the original investigators, focusing on a direct replication, being explicit about what is being tested, and so on. There is no current model of how to replicate a study and no consistency in the definition of replication (Schmidt, 2009). The R Project makes a huge gain in developing the model. In the process of clarification, it is likely that guidelines will emerge for reporting of research more generally. That is, we all ought to report or make available details that will allow easier replication and the R Project may help with that too.
5. The R Project emphasizes transparency. Individuals not involved in the project can examine the project's design and procedures, view replication materials from the various teams, look at reports or raw data from completed replications, and even join in to conduct a replication (see http://opensciencframework.org/project/EZcUj/). Making the procedures explicit and the materials available sets an excellent precedent for what we are trying to accomplish in science more generally. Findings and procedures are not secret or proprietary (with occasional exceptions on patented procedures or commercial materials). Also, so much of the research is funded through public dollars (e.g., federal and state grants or contracts given to universities). Sharing information and transparency are essential to science and not just something that is good to do. The R Project makes this view clear by how they are going about replication (transparency, involvement of many individuals, solicitation of others who might want to join).
6. Science has come under increased scrutiny in part because of fraud and fabrication that have occurred in biological, natural, and social sciences. Even though seemingly very infrequent, the circulation of information (e.g., Web, news media) is better or at more extensive than ever before and retractions (when authors and journals make some effort to "take back" and renounce what was published) are more visible and available too. And news media more routinely comment on scientific findings and reflect skepticism about replication and replicability of effects (e.g., Lehrer, 2010).

Sometimes provocative and inaccurate article headlines sensationalize more than help (e.g., Yong, 2012a, b). Even so, scrutiny and being watched more are all to the good. Within psychology, one fraud case, as I mentioned, was used in the media to indict social psychology, then all of psychology, then all social sciences, and "Dutch universities" (because the investigator was from one such university). In other words, fraud challenges the credulity of the enterprise at many levels. The R Project is a constructive effort to examine in an open and transparent way what the reliability of our findings are and answer questions we all have about how solid and reliable our findings. This is an issue for all science, and that psychology has taken such a leadership role is noteworthy and exemplary.[5]

Summary and Conclusions: Cautions, Negative Effects, and Replication

Three areas related to data interpretation were discussed in this chapter:

- Interpretation of the results of a research study
- Negative results
- Replication

In discussing the results of one's study, little inferential leaps often are made that can misrepresent or overinterpret what actually was found in the data analyses. Common examples from clinical research were mentioned such as stating something more than one is entitled to say based on statistical or clinical significance or based on the relation demonstrated in the study (e.g., use of the term "predictor"). The concepts discussed are basic, but it is surprising how often the investigator's interpretations of the data make little leaps that the findings do not warrant. Discussion of one's results requires and indeed demands going beyond the data and hence one has to be vigilant in one's own work and in the works of others. The issue is not trivial but has to do with the fundamentals of scientific epistemology—what do we know from this study, what can we say as a result?

Another topic critical to data interpretation is the notion of negative results, a concept that has come to mean

that no statistically significant differences were found in the experiment. The concept has received attention because the putative importance of a given study and its publishability often depend on whether statistically significant results are obtained. Unfortunately, such an emphasis has detracted considerably from other considerations, i.e., whether the conclusions can be accepted because of the theoretical or empirical importance of the question and quality of the research design, independently of statistical significance. Hence, methodologically weak studies with statistically significant results are more likely to be published and methodologically sound studies. Also, studies at least as sound as those published without statistically significant effects often go unpublished.

Related to the topic of "negative" results is the notion of replication or repetition of a previously conducted study. Replications can vary in similarity to the original experiment. Direct replications attempt to mimic the original experiment, systematic replications purposely attempt to vary the conditions of the original experiment, and extensions or conceptual replications move even further away from the original study. Replication research may lead to negative results, which can bring into question the basis for the results of the original experiment or the generality of the original findings. Replication research is exceedingly important because it is the most reliable test of whether the finding is veridical. The logic of statistical analyses suggests that occasionally statistical significance will be achieved even when there are no group differences in the population, i.e., findings significant by "chance" alone. Since these are likely to be published because of the bias for "positive" findings, there could well be a great many findings that would not stand up under any replication conditions. Thus, to distinguish those findings in the field that have a sound basis requires replication research. Also, there is the other side. Some "real effects" will lead to nonsignificant findings as a function of "chance" (but more likely weak power). Here the danger is not pursuing some negative effect that may be cast aside prematurely. Replications need not merely repeat a previous experiment but can address nuances of the original experiment as well as entirely new questions such as the conditions under which the relation is or is not likely to hold. Increased attention to the importance of negative findings and replication and concrete efforts to foster replications suggest changes that can improve science.

Critical Thinking Questions

1. Give two examples (hypothetical or real) where a negative result would be important.
2. What are three reasons why negative results might be due to poor methodology? (My dissertation cannot be used as part of this answer.)
3. Replication is so pivotal to science. Yet, if there is a publication bias, how can that interfere with the effectiveness of replication?

Chapter 15 Quiz: Cautions, Negative Effects, and Replication

Chapter 16
Ethical Issues and Guidelines for Research

Learning Objectives

16.1 Report ethical consideration in statistical studies

16.2 Recognize the importance of ethical considerations while handling nonhuman animal subjects during statistical tests

16.3 Describe some of the areas of statistical research where ethical issues are in the spotlight

16.4 Determine some critical ethical areas of statistical research

16.5 Evaluate the practice of informed consent of subjects as used in statistical experiments

16.6 Investigate ethical issues in applying intervention practices on subjects of statistical studies

16.7 Express the position that the law takes in guiding ethical statistical research

There are enormous responsibilities that are central to conducting research. For purposes of organizing the discussion, it is helpful to consider these under two broad rubrics: ethical issues and scientific integrity.

> *Ethical issues refer to the investigator's moral, professional, and legal responsibilities in relation to the care of research participants.*
>
> *Scientific integrity refers adherence to the standards, responsibilities, and obligations in conducting and reporting research.*
>
> To state these simply and to help remember the distinction, ethical issues focus primarily on the responsibilities of investigators in relation to participants; scientific integrity focuses primarily on responsibilities to science and the profession more broadly.

The issues overlap—both relate to science, have moral, professional, and legal issues, and reflect quite specific practices (e.g., what to do in a study) and character (e.g., integrity, honesty). Yet they have a different thrust. This chapter considers ethical issues.

Ethical issues are listed and treated toward the end of the text because they raise general issues that affect most if not all other facets of methodology. Yet, the issues and responsibilities are relevant at the very beginning of a study, in fact, at the stage where the studying is being designed and proposed. For example, when a study is being proposed and undergoes approval by a human subjects review committee (Institutional Review Board or IRB) at most universities, ethical issues receive attention at the outset. Approval of a proposal will require answering such questions as:

- Are the participants going to be subjected to any risk, including providing information that might be misused by the investigator?
- Are there any deceptive practices in the study?
- Is the potential knowledge yield of the study worth the potential risk that any participants might experience?

These are not all the questions, but they convey that ethical issues and responsibilities enter into the research process before the first subject is recruited. Evaluation and approval of the study and allowing it to go forward require that several ethical issues are satisfactorily addressed.

16.1: Background and Contexts

16.1 Report ethical consideration in statistical studies

This chapter focuses on critical ethical issues, research practices in which such issues are raised, and professional obligations associated with them.

16.2: Scope of Ethical Issues

16.2 Recognize the importance of ethical considerations while handling nonhuman animal subjects during statistical tests

In relation to ethical issues, the primary focus of the chapter will be on the conduct and reporting of research with humans. The focus demands recognition of the importance of what will not be covered.

First, ethical issues are critically important in the care and treatment of nonhuman animal subjects. There is enduring concern about animal rights, conditions for the care of animals, and "sacrificing" (killing) animals in research and classroom educational programs. Most of these and related issues have decades of history, but in many ways concern about the rights and protections of nonhuman animals has increased in the past decade or two. Among the reason is increased attention to the resemblance of many animals to humans (e.g., chimpanzees and dolphins) in relation to intelligence, consciousness, and understanding (broadly referred to as animal cognition) has led to a reduction in their use in research, a call for such reductions (Grimm, 2010; National Institutes of Health [NIH], 2013f). Many "animals" have quite sophisticated cognitive (e.g., intelligence, empathy, memory, planning) and affective processes (e.g., empathy, reaction to trauma) or engage in practices we ignorantly flaunted as unique to us. Just a teaser:

- Dolphins have names (sounds) for each other and call them out and remember names of old friends they have not seen for extended periods (Morell, 2013)
- Elephants without special training seem to respond to human cues (e.g., to select among alternatives) leading us to now consider elephants as a domesticated animal (like dogs) (Smet & Byrne, 2013). Also, elephants experience the stress of other elephants and use gentle touch to console each other (Morell, 2014)
- Monkeys and chimpanzees engage in the equivalent of war: They unite as a group and attack adjacent groups (Aureli, Schaffner, Verpooten, Slater, & Ramos-Fernandez, 2006)
- Bacteria communicate through biochemical signals with one-another uniquely (to their own kind) and to other bacteria and base their action on when the population of other bacteria reaches a particular number, a phenomenon called quorum sensing (e.g., Rumbaugh, 2011)

Probably no one believes bacterial communication is like ours or that one bacterium goes to another bacterium with a relationship issue and says, "We have to talk." Yet the broader point is the main one—the capabilities of nonhuman animals in communication, cognition, affect, and action have been enlightening and humbling. That work has contributed to increased concern about using nonhuman animals as subjects but certainly using them in other ways too (e.g., in captivity, as sources of food). For this chapter, I will defer to other sources for considerations of critical ethical issues raised in the care, protection, and use of nonhuman animals (e.g., Sandøe, 2013; Sikes & Gannon, 2011).

Second, ethical issues are critically important in clinical psychology in the practice of the profession, including primarily the delivery of psychological services, assessment, and consultation.

> The practice of the profession is governed by a set of ethical codes and guidelines (e.g., licensing and certification requirements for certain types of practice, restrictions on the personal relationships one has with one's clients).

Needless to say, the ethical issues and guidelines that govern the application of psychology are on par with those related to research, but are beyond the scope of a text on research methods and clinical psychological science.

This chapter will not address ethical issues specifically related to nonhuman animal research or the practice of clinical psychology. Ethical codes for the profession of psychology (e.g., American Psychological Association [APA], 2010a) address all of these issues. The core issues that relate to the conduct of research are sufficiently weighty to narrow the focus of this chapter.

16.3: Inherent Roles of Values and Ethics in Research

16.3 Describe some of the areas of statistical research where ethical issues are in the spotlight

Occasionally a salient view has characterized science as the pursuit of knowledge that is largely value free and ethically neutral. The stereotypic image is lovely—a scientist in an out of fashion ill-fitting dress or rumpled slept-in-look khaki pants and with at least one article of clothing in the overall outfit unbuttoned somewhere. The person is huddled near a piece of equipment or staring at a computer screen and working intently on a scientific problem. The image strongly implies that one is in a lab and cocooned from controversies and ethical and social issues. The search for knowledge and the methods used in that search were once considered to be ethically neutral. That is, findings from research may be misused, but science itself is sort of above it all. The neutrality view is long gone because it is so easily challenged. A few wars ago (World War II),

illusions of neutrality were clearly erased. The development and deployment of weapons for war (e.g., atomic bomb, radar, rocket launchers) are feats of science and technology, and scientists have played a central role. Their invention often is in the service of governmental ends, but even when they are not, they can be used for ends that raise all sorts of ethical issues (e.g., cloning, making new or hybrid species, robots for use in war).

Many topics of contemporary research (e.g., developing treatments that can be used throughout the world, genetic engineering of foods, animals, and people; use of control groups that might withhold promising interventions from life-threatening conditions) have made it more clear than ever before that science is laden with values, whether this reflects the focus of research, the conclusions one reaches, implications that one draws, or how the science is translated in the law and policy (e.g., Ehrlich & Ehrlich, 2008). The values are reflected in decisions what to study, what to fund in research, conflict of interest, and abuses in advocating the use of findings. This in turn has led to the view, advocated here, that ethics is not an ancillary part of research but rather deeply intertwined with the entire enterprise.

16.3.1: Values and Decisions in Research

> *Increased attention to ethical issues has resulted from research in public view in which there have been clashes between the immediate benefits to individual participants and knowledge or long-term benefits to yet-to-be-identified persons.*

Also, attention has been mobilized in response to dramatic abuses and research practices (e.g., fraud, conflict of interest, death of subjects during research). Again many of the issues are not new. Yet, the global scale of science (e.g., multisite studies across many continents and countries), the visibility of science in the media, and increased activism and scrutiny of science by the public (e.g., through the Web) have placed a microscope on precisely what is going on and why and whether it should.

The value issues are easily illustrated by examples from the treatment of HIV. In more than one study, HIV patients have been assigned (randomly) to a control (no-treatment) group or to a placebo group. Some of the research has been conducted in developing countries (e.g., South Africa, Thailand, Ivory Coast). Some researchers have argued that standard treatment in the United States, already known to help HIV, is not likely to be the standard treatment in developing countries because it is too expensive or simply unavailable. Thus, "no treatment" is the standard care in such countries. This rationale has been used to justify including no-treatment control or more often the use of placebo-control conditions. Needless to say, this position is quite controversial and has generated extensive discussion, led to guidelines for research, and raised ethical issues of research on a worldwide scale (e.g., Holland, 2012). More generally should a viable treatment ever be withheld from a person in need, whether or not they agree to that? This is an ethical as well as a research issue.

16.3.2: Relevance to Psychological Research

Many of the issues (e.g., withholding medical treatment) seem only tangentially relevant to psychology. Yet, they are directly relevant in two ways:

1. Many issues that might seem unique to biological sciences and medicine are central to psychological research (e.g., collecting genetic information, deciding what control groups should be in clinical trials of psychosocial treatments in developing countries). Research now is more collaborative and multidisciplinary than ever before, and major research projects often involve multiple settings, countries, and disciplines and with that a richer set of ethical and legal issues to consider.
2. Even when a particular line of research may not seem relevant to psychology, that research may generate guidelines, laws, and federal and university review criteria that govern all research and place into the limelight subject rights and privileges more generally. I mention guidelines below from many agencies, and these guidelines apply to all research whether or not the issues that prompted the research emanated from psychological studies.

Focusing specifically on psychological research, it is easy to identify the situations that routinely raise ethical issues. Actually, the "basics" of psychological experimentation raise issues:

1. Experiments require manipulation of variables, which often may subject participants to experiences that are undesirable or even potentially harmful, such as stress, failure, frustration, and doubts about themselves.
2. Implementing most experimental manipulations requires withholding information from the subject. The experimental question may address how subjects respond without being forewarned about exactly what will happen or the overall purpose.
3. Experimentation requires assessment or observation of the subject. Many dependent measures of interest pertain to areas that subjects may consider private, such as views about themselves, beliefs about important social or political issues, and signs of adjustment or maladjustment.

   One of the most private sources of information (believe it or not) in clinic and community samples is personal or family income.

This seemingly simple descriptive information from the sample may raise concerns that the information might be publicly disclosed and have untoward implications (e.g., for collection of social assistance, payment of income taxes, custody support, loss of social assistance). In my own clinic, it is much easier to find out information about who is abusing whom in the home, whether one or more parents are taking what illicit substance, and who has a criminal record than it is to obtain accurate information about income. Clearly, issues pertaining to invasion of privacy and violation of confidentiality are raised by assessment and even by assessment that may not appear to be very weighty.

Finally, much of the data collected are obtained online as measures are administered via the Web. There are huge concerns of privacy and hackers obtaining personal information. The use and abuse of online information and the many ways that information is unwittingly shared (e.g., loss of laptops, transferring or peeking of files in unauthorized ways by employees, successful hacking) have added further concern about individual rights and protections.

16.3.3: Power Difference of Investigator and Participant

Ethical issues also are raised by the relationship between the investigator or experimenter and the subject. The difference in the power and status of the experimenter and access to information (e.g., about procedures, hypotheses, goals, and likely outcomes) allows for potential abuses of the rights of the individual participant. The power differential means the subject is not an equal in making informed choices about participation in the study. For example, subjects might be embarrassed, concerned, or feel foolish given the manipulation and its focus if they knew all there was about the study. The "titles" of studies (evaluation of cognitive processes) often obscure the actual focus (self-regulation, altruism under stress). Research participants, particularly in clinical research, often are disadvantaged or dependent by virtue of their:

- Age
- Physical and mental condition
- Captive status
- Educational level
- Political and economic position

For example, samples in clinical research may include children and adolescents, psychiatric patients, the elderly, victims of domestic violence, prisoners, person in need of (and maybe desperate for) treatment, and individuals who cannot pay for services in the usual way because they do not have insurance or health coverage. These subjects might be more readily induced into research and have, or at least feel they have, relatively little freedom to refuse or discontinue participation in light of their options.

The status of the investigator is sustained by several factors. The investigator structures the situation in which the subject participates. He or she is seen as an expert and as justified in determining the conditions for performance. The legitimacy, prestige, and importance of scientific research all place subjects in an inferior position.

> Although subjects can withdraw from the research, this may not be seen as a realistic or very likely option, given the status differential.

Subjects may see themselves as lacking both the capacity and right to question the investigator and what is being done. Subjects are at a disadvantage in terms of the information about the experiment at their disposal, the risks that are taken, and the limited means for counteracting objectionable aspects of the treatment.

In current research, guidelines to protect subjects require that investigators inform participants of the goals of the study and any risks and benefits that might accrue.

> Written consent is obtained to confirm that subjects understand the study and their rights not to participate and to withdraw.

Both legal codes (e.g., for universities receiving any federal funds) and ethical codes (from professional organizations) guide the process of disclosing information to and obtaining consent from the subjects. The codes and practices that follow from them are designed to ensure that the rights of the individual subject are protected and are given the highest priority in research. At the same time, consent procedures are designed as well to protect research institutions that must document that appropriate protections were taken, are in place, and have been documented. In the background are overarching concerns (e.g., litigation, suspension of funds from funding agencies) against which institutions wish to protect. If anything goes wrong (e.g., unexpected side effects, death, really being upset), the university wants to be in a position of saying and showing, "we did everything we were supposed to do." Consent procedures are part of that.

16.4: Critical Issues in Research

16.4 Determine some critical ethical areas of statistical research

Although many ethical issues can be identified, a few seem particularly salient:

- Using deception in experiments
- Informing participants about the deception after the experiment is completed

- Invading the subject's privacy
- Obtaining informed consent

Whether the research is laboratory-based (e.g., college students completing a cognitive task, community members participating in a study of stress) or clinic-based (e.g., intervention research with clinic samples), these issues can easily emerge.

16.4.1: Deception

Deception may take many different forms and can refer to entirely misrepresenting the nature of an experiment at one extreme to being ambiguous about the experiment or not specifying all or many important details at the other extreme.

The extent to which these various active (e.g., misrepresentation) or passive (e.g., failure to mention specific details) forms of deception are objectionable in part depends upon the situations in which they are used and the effects they are likely to have on the participants.

Misleading the participant may not be very objectionable in many experimental arrangements. For example, when participants perform a memory task involving lists of words or syllables, they may be told initially that the purpose is to measure the way in which individuals memorize words. In fact, the purpose may be to assess the accuracy of recall as a function of the way in which the words are presented. In this situation, there seems to be little potential harm to the participants or their evaluations of themselves. Alternatively, participants may be placed under psychological stress or led to raise important questions about themselves. For example, participants may be misled to believe they are very competitive, have latent sexual problems, or are odd in the way they think. The goal may be to evaluate these induced states on some other area of functioning.

An illustration of deception in psychological research is provided by the well-known experiments of Stanley Milgram (1933–1984), who conducted research on obedience to authority. The study began in 1961 with findings first published in 1963. The years help establish the context for the research, which was the Nazi war crimes and Nuremberg War Criminal trials of World War II. Those crimes and the trials revealed that many who were involved in acts of genocide and other horrible acts they were just following orders, i.e., were being obedient.

Milgram conducted laboratory experiments in which subjects were recruited to evaluate the extent to which individuals would engage in acts that were cruel in response or at least analogous to following orders. They did not know they were the subjects and were called "teachers" who were going to help others who were "learners." The teachers were given a fictitious story that the experiment was intended to explore the effects of punishment on learning. The learners actually were confederates, i.e., actors working as part of the experiment and it was the teachers who were really the subjects. The teachers were asked and encouraged to increase shocks, including intense shocks to punish the learner. In fact, no actual shocks were given but the teachers (i.e., subjects) did not know this. The results essentially showed that many—in fact over 60%—of the teachers (subjects) administered high doses of the shock (see Milgram, 1963, 1974). The results convey the high level of obedience to authority. There are many interesting features of this research that can easily be found on the Web (e.g., search "Stanley Milgram" on most search engines) and other sources (e.g., Perry, 2012; Russell, 2011). As one can discern from these references, Milgram's work and the issues they raise are still very much alive.

Attention to that work increased in the context of torture of prisoners following up on terrorist attacks in the United States in 2001. Individuals who were accused of terrorism were often tortured by soldiers. The soldiers were placed in extraordinary situations (prisons guarding terrorist prisoners) and engaged in cruel behavior suggesting again that humans can readily be pushed to engage in behavior they would otherwise not do (see Zimbardo, 2007).

The issues of torture and public outcry about the use of techniques by soldiers who guarded the prisoners brought back to public discussion the Zimbardo experiments on prisoners and guards conducted as a university psychology experiment (see Zimbardo & Cross, 1971). As is well known some students were randomly assigned to be guards and others as prisoners in a mock prison constructed in the psychology building. The goal was to observe the interaction of the two groups. The guards ended up being harsh and abusive after assuming their roles, such that the experiment had to be terminated in 6 days. The conclusions focused on the roles that generate what people do even without specific orders or requests to be obedient (as in the Milgram study). Obedience and cruelty remain critical topics for psychological research (Burger, 2009; Fast, Halevy, & Galinsky, 2012; Slater et al., 2006). The ethical issues, controversies, and practices raised by the Milgram and Zimbardo studies continue to receive attention (e.g., Johnson, 2013; Sontag, 2012). The public and research community is more sensitized than ever before both to substantive issues (obedience, cruelty) and ethical issues related to the conduct of research.

Deception still is possible in contemporary studies of psychology, but the bar for allowing that is much higher than it was decades ago in the "classic" studies (Milgram, Zimbardo) of yesteryear. Also, rarely are the personal and social issues of everyday experimentation at the level of the moral dilemmas raised by those studies. Even so, the broad issue remains the same, namely, deciding whether deception is justified in a given experiment and whether the possible risks to the participant outweigh the potential

benefits in the knowledge the study is designed to yield. Both the risks to the participant and potential benefits usually are a matter of surmise, so the decision is not at all straightforward. The dilemma is particularly difficult because the risks to the individual subject are weighed against the benefits to society. In most psychological experiments, the benefits of the research are not likely to accrue to the participant directly. (In so many cases, the benefits to society are not so clear either, or at least there is rarely even a hint of follow-up evidence that evaluates when or whether an experiment or line of work has benefitted anyone.) Weighing potential benefits to society against potential risks to the individual subject is difficult. The safest way to proceed is to minimize or eliminate risk to the subject by not using active forms of deception.

> The potential harm that deception may cause for the individual subject certainly is a major ethical objection to its use.

Moreover, aside from its direct harmful consequences, the act of deception has been objected to because it violates a value of honesty between individuals, in this case the investigator and subject. Investigators engage in deceptive practices that would not be condoned outside of the experimental setting because they violate the basic rights of individuals. Thus, deception fosters a type of behavior that is objected to on its own grounds independently of its other consequences. Alternatively, it may not be the deceptive behaviors in which investigators may engage as much as the context in which these behaviors occur. Many forms of deception occur in routinely everyday life (e.g., Santa Claus, Tooth Fairies just for starters) and individuals may love some of these (e.g., surprise proposals for marriage) but despise others (e.g., surprise parties). The problem with forms of deception and surprises in an experiment is that the professional context of an experiment may lead people to expect full disclosure, candor, and respect for individual rights. This shift in term from "subject" to "participant" in part is to recognize that people are not just subjects (i.e., are not just subjugated) but rather active contributors to the process and knowledge. This recognition reflects greater consideration of their right to more equitable treatment.

16.4.2: Further Considerations Regarding Deception

Actually, deception in clinical, counseling, and related areas of research rarely involves efforts to mislead subjects. Rather than presenting misinformation, it is likely that some information will be withheld. How much to withhold? Should the participants be aware or fully aware of the purpose and procedures of the experiment? Ideally, investigators would fully disclose all available information about what will take place. Complete disclosure would entail conveying to subjects the nature of all of the procedures, even those to which the subjects in any particular condition will not be exposed, and revealing the investigator's view and expectations about what the results might yield. In most psychological experiments with human subjects, full disclosure of available information may not be realistic. If the subject knows the purpose, hypotheses, and procedures, this information could influence or alter the results. This raises construct and external validity issues.

> *Construct validity* is raised because the effects may not be due to the manipulation alone but the manipulation combined with knowledge of what is expected (the hypotheses).
>
> *External validity* is a threat because the findings may only apply to individuals who are aware of the hypotheses (reactivity of arrangements).

For example, we want to know how people regulate (control) anger to understand fundamental features of that process and perhaps ultimately be able to help people control that anger (e.g., domestic violence, road rage, school shootings). The basic work, or so it would seem, ought to not say to participants ("We want to see if and how you control anger in this laboratory setup and expect that you will be able to control yourself in one situation but probably not another").

The concern that disclosure can influence the results has been known for a long time. Indeed, with disclosure first became mandated—over 4 decades ago—a simply laboratory study revealed the impact on findings. In this study, college students participated in a verbal conditioning experiment in which their selection of pronouns in a sentence-construction task was reinforced by the experimenter by saying "good" or "okay" (Resnick & Schwartz, 1973). Some subjects (informed group) were told that the purpose was to increase their use of "I" and "we" pronouns in order to determine whether telling subjects the purpose of the experiment affected the results. These subjects were told the true purpose (i.e., to evaluate the effects of full disclosure). Other subjects (uninformed group) were told that the experiment was designed to study verbal communication. They were not informed of the real purpose of the experiment. Subjects in both groups constructed sentences and received approval when "I" or "we" pronouns were used in the sentences. As expected, the uninformed subjects increased in their use of the target pronouns that were reinforced over their base rates in a practice (nonreinforced) period, a finding shown many times in prior research. In contrast, the informed subjects *decreased* in their use of target pronouns relative to their initial practice rates. Thus, disclosing information about the purposes of the experiment completely changed the findings. This is not a shock in terms of psychological processes.

Efforts to control or manipulate behavior when recognized as such can readily lead to opposite behavior and have been studied in different ways in psychology (e.g., reactance, countercontrol).

The results suggest that informing subjects about the purposes and expected results of an experiment might dictate the specific relation that is obtained between the independent and dependent variable. Of course, one might view the results in another way, namely, that not telling subjects about the experiment dictates a specific relation as well and one that is not more or less "real" or informative than results obtained under informed circumstances. Yet, a major goal of psychology is to study behavior and to extrapolate findings to those circumstances in which individuals normally behave. That is, most of the theories we develop about human functioning are not intended to account for phenomena only evident in the laboratory situation or under obtrusive and reactive arrangements. Thus, investigators wish to understand how subjects respond to events when subjects are not forewarned about their anticipated effects and the purpose of exposure to these events.

> Although investigators, in principle, would like to avoid deception, it may be necessary in some form to understand certain cognitive, affective, and behavioral processes.

There is another situation that involves deception related to intervention research (e.g., in medicine and clinical psychology), namely, placebo effects. Placebo effects (changes in participant behavior as a function of expectations) are "real" and are based on participants believing they are receiving a real treatment. For example, sham (fake) surgery and medications often are control conditions in clinical trials and in fact greatly benefit the patients often as much as the true intervention (e.g., for knee injury, pain control) (Brim & Miller, 2013; Moseley et al., 2002). Stated another way, patients who are "deceived" by not being informed that they are receiving a placebo can benefit greatly! Yet, this is not always the case. For example, in some cases (e.g., surgery to infuse a gene transfer in the brain for Parkinson's disease) patients improved with sham surgery but not as much as the full surgical intervention (e.g., LeWitt et al., 2011). Placebo procedures of all kinds can benefit patients but not always to the same extent as the veridical treatment. To disclose or not disclose this was a placebo condition in advance is a matter of debate and differing positions. This is a special case because individuals can benefit from the deception and receive an intervention with fewer risks (side effects, complications of surgery) than the intervention may provide. I am not advocating a position but conveying that "to be [deceptive] or not to be [deceptive]" drawing on Hamlet's dilemma in Shakespeare, is NOT the question. How, when, how much, why, and for what population of participants (e.g., children, college students, clinical patients) all contribute to the appropriateness or suitability of not providing full disclosure.

In general, guidelines for informing subjects are dictated by law and by ethical principles that govern research and informed consent (e.g., United States Department of Health and Human Services [DHHS], 2009a; APA, 2010a). I shall say more about guidelines later, but it is important to note here that guidelines do not require elaborating all of the views, hypothesis, expectations, and related possibilities to the subjects. Thus, some information invariably is withheld. Of special concern in relation to deception are active efforts to mislead subjects. Such efforts are rare in clinical, counseling, and educational research. Research proposals that include efforts to mislead subjects must establish that deception is essential to achieve the research goals. Moreover, if deception is essential, the research now must specify special procedures that will be provided to reduce any lingering effects of the deceptive experience once the experiment is finished.

To establish that deception is necessary, at least three criteria must be met:

1. The investigator who designs the experiment must make the case to others that deception is justified given the importance of the information that is to be revealed by the experiment. An investigator may not be the best judge because of his or her investment in the research. Hence, review committees involving individuals from different fields of inquiry ordinarily examine whether the proposed procedures are justifiable. The committees, formally developed in most universities and institutions where research is conducted, follow guidelines for evaluating research and for protecting subjects, as discussed later in the chapter.
2. If there is any deception in the planned experiment, there must be assurances that less deceptive or nondeceptive methods of investigation could not be used to obtain the information. This too is difficult to assess because whether similar methods would produce the information proposed in an experiment that uses deception is entirely an empirical matter.

   > Researchers genuinely disagree about the extent to which deception is essential. Even so, an investigator is required to make a well-reasoned case.

3. The aversiveness of the deception itself bears strongly on the justification of the study. The aversiveness refers

to the procedures, degree or type of deception, and the potential for and magnitude of harmful effects. Deceptions vary markedly in degree, although ethical discussions usually focus on cases where subjects are grossly misled about their own abilities or personal characteristics. Will there be lingering aftereffects, and will a person be made to experience emotional or physical pain as a result of the deception, as distinguished from the procedures of the experimental manipulation? These help define aversiveness.

Research begins with the view that individual rights are to be protected. Investigators are to disclose to the extent possible details of the design, purposes, risks, benefits, and costs (e.g., monetary or other). The purpose is to permit the subject to make an informed decision regarding participation. If deception is to be used, either by withholding critical information or by misrepresenting the study, the onus is on the investigator to show cause at the research proposal stage that this is essential for the necessary benefits of the research. Unless the case can be made to review committees that evaluate such proposals, the work may not be permitted.

In many cases, even if deception seems necessary, the investigator's creativity and methodological and statistical skills can provide a path to obtain the information without deception.

It is useful to begin with the premise there may be no need to deceive subjects. Alternative experimental procedures may address whether deception is necessary. For example, the investigator may present to different groups varying degrees of information and see if this affects the findings. Alternatively, perhaps the methods used to evaluate demand characteristics such as the preinquiry or use of simulators can be explored to evaluate if subjects would be likely to perform differently under different conditions of disclosure. The absence of differences between groups studied in this way is consistent with the view that deception may not be critical to the research findings and methods of study in the area of work. These alternatives are not perfect in providing unambiguous answers that might be obtained with deception. These options may begin to make the case to the investigator that deception will be needed to pursue a particular question.

16.4.3: Debriefing

If there is any deception in the experiment or if crucial information is withheld, the experimenter should describe the true nature of the experiment after the subject is run.

Providing a description of the experiment and its purposes is referred to as debriefing.

The purposes of debriefing are twofold:

1. The goal is to counteract or minimize any negative effects that the experiment may have had.

   By debriefing, the experimenter hopes the subjects will not leave the experiment with any greater anxiety, discomfort, or lowered self-esteem than when they arrived.

2. A goal of debriefing is educative. Debriefing ought to convey the value and goals of the research, why the information is or might be important, and the participant has contributed to research (e.g., Moyer & Franklin, 2011). Discussions of debriefing understandably emphasize overcoming the deleterious effects of deception. Of all things, we do not research to harm participants in any way and deception could easily do that, considering harm broadly.

The manner in which debriefing is conducted and the information conveyed to the subject vary enormously among experiments. Typically, subjects meet with the experimenter immediately after completing the experimental tasks. The experimenter may inform the subject what the experiment was "really" about and explain the reasons that the stated purpose did not convey this. The importance of debriefing varies with the type of experiment and the nature of the deception. As part of the experiment, subjects may have been told that they have tendencies toward mental illness or an early grave. In such situations, subjects obviously ought to be told that the information was not accurate and that they are "normal." Presumably such information will be a great relief to the subjects. On the other hand, subjects may be distressed that they were exposed to such a deception or that they were "gullible" enough to believe it.

16.4.4: Further Considerations Regarding Debriefing

Table 16.1 notes the key elements of debriefing. These are provided in written form to the subject at the end of the experiment. The specific form and procedure must be approved by Institutional Review Boards before proceeding with the experiment. The table also provides a sample paragraph for debriefing to convey the content, style, and brevity that can satisfy. Because debriefing is supposed to be educational, some ingredients are included to explain the rationale for the study, prior work leading to the study, and then some background readings.

Debriefing has been assumed to be a procedure that resolves the potentially harmful effects of deception. Yet debriefing subjects by providing full information about the

Table 16.1: Debriefing: Elements and Sample Form

Elements of a Debriefing Form
• Title of the Project
• Statement of the Purpose of the Study
• Brief Background of Prior Work Leading to this Study
• Specific Hypotheses and Variables Studied
• Statement of what Part(s) was deceptive and why Deception was used
• Reminder that Data will be Confidential
• Mention that participants can Receive the Final Report of the study
• Provide contact information of the Investigator and Institutional Review Board Chair
• Provide some References for further reading

Sample Language Explaining Deception and Selected other Elements*

"In this study, we told you that you would receive a blue sticker and then we would ask you to report about how you felt about the sticker. Instead, we gave you a red sticker and told you that your friend took the last blue sticker. However, this was not true; your friend didn't take the last blue sticker. We did not tell you the full nature of the experiment because we wanted to gauge your honest reaction to the news that your friend took your sticker. Stickers, and the way that friends react to them, provide interesting insights into interpersonal relationships. In previous studies . . ., blue was found to be particularly desirable, thus it was chosen in order to evoke a stronger response. We are interested in learning if there is a correlation between individuals who are more capable of negotiating the lack of a blue sticker and their ability to maintain a friendship. Please know that your friend was not involved in this study and had nothing to do with the blue sticker. It is important that you do not let this incident become an issue in your relationship. If you feel that you didn't negotiate the loss of a sticker in a positive way, this may be an opportunity to evaluate your friendship and learn what you can do to better handle this situation should it arise. The "Sticker Group" is an informal friendship counseling group available for . . . students; for more information, see their website: . . . If you have further concerns, please contact the researcher (name, contact information) to discuss any questions about the research. If you have concerns about the way you were treated as a participant in this study, please contact the . . . Chair, Institutional Review Board (full name, address, and phone number)."

The Sample Paragraph was adapted from University of Virginia (2013, HYPERLINK "http://www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html" www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html)

University of Virginia (2013); Institutional Review Board for Social and Behavioral Sciences. Sample briefing statement. Retrieved from HYPERLINK "http://www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html" www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html)

deception may not necessarily erase the false impressions established during the experiment. This is roughly analogous to someone lying to you and much later saying he or she had a good reason at the time or did not really mean it. The initial lie may place small stain in the fabric of one's relationship that is not completely erased by stain-removing comments.

> The fact that the effects of deception may linger even after debriefing provides us with reason for further caution.

In the sample paragraph in Table 16.1, the participants were given recourse to attend counseling if they had further issues and also encouraged to ponder their relationships. These points in the paragraph very much raise the prospect that being deceived may not be adequately resolved by merely explaining what was done and why. Understandably, if deception is to be considered, it must be quite clearly justified to assure the risks to individual rights and integrity.

The timing of debriefing may be important and relevant to its success in countering any lingering effects. Sometimes experimenters wait until all subjects complete the experiment and contact subjects with a printed handout or class announcement. The reason for this is that information provided early in the experiment can filter to other subjects before they serve in the study (Edlund, Sagarin, Skowronski, Johnson, & Kutter, 2009). However, delayed debriefing may not be as effective as immediate debriefing. If subjects are potentially harmed by the deception, the experimenter's obligation is to debrief as soon and as effectively as possible.

Occasionally, not debriefing participants is argued as ethically acceptable. Among the situations in which foregoing debriefing is considered reasonable are those in which debriefing is not very practical (e.g., reaching all individuals after the study is completed), the deception seems innocuous, and reasonable people would not object to the deception for purposes of research (e.g., Sommers & Miller, 2013). The use of deception and debriefing are judgment calls, and it is important to not rely solely on one's own judgment where there is a vested interest about the value of a project.

> Professional guidelines, input from colleagues, and formal evaluation by Institutional Review Boards all are resources to aid in decision making.

16.4.5: Invasion of Privacy

> *Invasion of privacy refers to seeking or obtaining information of a personal nature that intrudes upon what individuals view as private.*

In research projects, information may be sought on such topics—use of drugs, sexual beliefs and behaviors, health, income, and political views. People often are hesitant to share information on such topics and respond in ways to provide socially desirable responses and to limit disclosing information (Bansal, Zahedi, & Gefen, 2010; Tourangeau & Yan, 2007). Beyond research, other sources solicit information from individuals, including credit bureaus, investigative and sales agencies, and potential employers.

The use of tests that measure psychopathology and personality also raises concerns over invasion of privacy. Test results can reflect directly upon an individual's psychological status, adjustment, and beliefs and uncover personal characteristics that the subject might regard as private. Moreover, the information obtained through

psychological testing might be potentially damaging if made public.

> The threat of personality testing to the invasion of privacy has been a topic of considerable concern. One reason is that measures of psychopathology and personality have been used routinely to screen potential employees in government, business, and industry.

Many of the questions asked of prospective employees seemed to be of a personal nature and not clearly related to the tasks for which individuals were being selected. Some of the items may even focus on illegal behavior (e.g., use of some drugs, admission to criminal activity such as abusing one's child or partner).

Advances in technology and social media have greatly expanded the scope of concern and threat to privacy invasion. The use of information that is posited through social media (e.g., Facebook) represents one facet that can be considered invasion of privacy, at least if the individual is seen, hacked, or otherwise used by individuals who are not the intended audience. More broadly, use of the Internet raises multiple opportunities for invasion of privacy by tracking all sorts of information (e.g., one's friends to whom we are linked) and unbeknownst to us connecting our activities (e.g., what sites we use, with our Internet address).[1] Also, one's own personal computer can be "hacked" to find out more personal information (e.g., social security numbers, credit card numbers, bank accounts, and maiden names of various maidens in one's life).

> Most hospitals and clinical practices associated with them are moving or have already moved to electronic records rather than printed files housed in some file cabinet.

Now health information is more readily integrated and more readily available to all of one's doctors. Hacking those secure systems too is a matter of time. Also, the issue about secure electronic information is not about hacking or obtaining information illegally. Much information can be obtained (e.g., all cell phone calls, all e-mails, all Web site visits) legally in some states where the information can be easily subpoenaed by attorneys, for example, who make the case that the information is pertinent to something on which they are working.

Finally worth mentioning is burgeoning work on genetics and mapping of the genome. Increasingly psychological research focuses on facets of the individual genome, and data are collected to characterize a given sample. For example, gene association studies have focused on predicting aggression in patient and nonpatient populations (e.g., Vassos, Collier, & Fazel, 2013; Zalsman et al., 2011).

> Protection of genetic information is important now that the genome can be used to identify propensities for physical and mental disorders.

It is easy to imagine how that information could be abused (e.g., by prospective employers, insurance companies, attorneys building a case against someone in a heated divorce, and in the frequent negative campaigning for political office). This is a new kind of "identity theft" beyond stealing credit cards. Advances in research often require advances in the means of protecting the information. In short, invasion of privacy has a variety of threats and even more than in the past given social media, online activities, and electronic records of medical and biological (e.g., genetic) characteristics.

In psychological research, the major issues regarding invasion of privacy pertain to how the information from subjects is obtained and used. Ordinarily, information provided in experiments must to be provided willingly. Obviously, there are many kinds of research where consent of the individual is neither possible (e.g., in cases of severe psychiatric or neurological impairment) nor especially crucial (e.g., in the case of studying archival records for groups of unidentifiable subjects). Psychological research increasingly is conducted over the Web, and subjects provide their answers to measures that appear on their computer screen. The subjects may identify themselves by password, e-mail address, or by subject and demographic variables (e.g., age, sex). In such research, there may be assurances that the information will remain confidential. Subjects are appropriately wary because it is usually quite easy to trace information to a particular computer, unless one goes to special lengths to erase one's tracks. Also, businesses that claim the information (e.g., credit card number, e-mail, social security) provided by customers over the Web is confidential already have a few well-publicized lapses in security. We all recognize that information over the Web is not private and in principle assurances cannot be given that the information is completely confidential. This concern is more likely to emerge as diagnosis, assessment, and psychotherapy are conducted over the Web.

16.4.6: Sources of Protection

Three related conditions are designed to protect the subject's right to privacy in research are:

- Anonymity
- Confidentiality
- Protected access to one's records

> *Anonymity refers to ensuring that the identity of the subjects and their individual performance are not revealed*. Participants who agree to provide information must be assured that their responses are anonymous.

Anonymity can be assured at the point of obtaining the data as, for example, when subjects are instructed not to identify themselves on an answer sheet or computer-delivered

measure. Perhaps the study only requires basic subject and demographic information (e.g., age, sex, ethnicity, marital status, income) and the name is not needed at all. Much of the research conducted online (e.g., surveys, MTurk, Qualtrics) is of this type.[2] If names are collected for some reason, the names of participants are separated from the measures to eliminate any association with the scores on the measures. Typically, participants are given a code number in the database and in any files (physical or electronic) of the data that are kept separate from that, if kept at all. Only the investigator has the key to the code in which the participant's subject number and name are kept. In short, in most research, anonymity is assured by not seeking identifying information (name, date of birth) to begin with or when such information is obtained coding it immediately so that measures and their scores whether in electronic or paper form are disassociated with participants' names.

Once the information is obtained, participants must be assured that their performance is confidential.

> *Confidentiality means that the information will not be disclosed to a third party without the awareness and consent of the participant*.

Of course, if the participants cannot be identified (anonymity), then confidentiality usually is assured (but not always as we shall see). Conceivably, situations might arise where confidentiality is violated as, for example, when the information might conceal some clear and imminent danger to an individual or society. For example, clinical psychologists are involved with research on the evaluation, treatment, and prevention of AIDS. *Confidentiality* about who is participating in the research and about test results for infection is obviously important. The information, if inadvertently made available, can serve to stigmatize research participants and subject them to discrimination in everyday life (e.g., employment, housing). In some cases, information may emerge that has to be reported even though subjects would choose not to have the information revealed. For example, a study on parenting or child–parent interaction may reveal that there is child or sexual abuse in the home. In most states, this has to be reported to child services as a matter of law. In any case, confidentiality invariably has to be assured in research baring situations where there might be danger or a violation of law by withholding the information.

> Finally, protected access and privacy of patient records is yet another form of protection from invasion of privacy.

Beginning in 1996, with subsequent revisions, in the United States a federal law was passed called the Health Insurance Portability and Accountability Act (referred to as HIPAA—pronounced—Hip-uh) (www.hhs.gov/ocr/privacy/hipaa/understanding/index.html). The major goal of HIPAA is to ensure privacy of client health information. Privacy refers to an individual's right to control access to and disclosure of health information provided by the patient but also notes or observations by health practitioners that enter the patient's record. "Health information" is defined broadly and includes physical and mental health, psychological problems, and special services of other types (e.g., special education programming). Institutions that deal with patients or private information (e.g., mental, physical health) must identify a specific person as a privacy officer who oversees enactment of HIPAA guidelines. Also, specific procedures are checked to ensure that information remains private (e.g., all records electronic or other) and is not available to unauthorized personnel without permission of the client.

Among the reason is that many persons may have special interest in the information (clinical records) obtained as part of the research (e.g., employers, relatives, school administrators, attorneys involved in divorce, custody, and estate disputes) and that the nature of the information (e.g., measures of adjustment, psychopathology) may be potentially damaging if misinterpreted or misused. HIPAA is designed to:

- Protect participants
- Give them rights over their health information
- Recourse if their rights are violated

Violations of HIPAA by unauthorized use or disclosure of information have penalties associated with them (small or huge fines, prison terms).

> In research, data have to be "de-identified," the term used to note that the information (e.g., responses to questionnaires, disclosure of personal information) is coded in such a way that the information cannot be connected with specific client names.

In addition, several safeguards area required when private information is scored on a computer (e.g., special encryption software, use of university rather than personal computers). Clearly, HIPAA provides additional protections beyond anonymity and confidentiality we have already discussed.

16.4.7: Special Circumstances and Cases

So much of psychological research is based on using college students participating in a laboratory experiment or subjects from online sources where identity of the participant is not revealed. Those situations rarely require attention to special protections for privacy. Yet, invasion of privacy enters into many different areas in clinical research. For example, privacy is an important issue in writing the results of research investigations and treatment applications. Clinical research

reports occasionally are prepared for publication where an *individual case* is involved. In these instances, efforts to maintain confidentiality require the investigator to disguise the ancillary information about the client in such a way that his or her identity could not be recognized.

> Typically, pseudonyms and ancillary facts (e.g., slight changes in age or other demographic variables) are used when a case is described in published form.

Yet for many case reports, a change in the name and minor other changes may not protect the subject's confidentiality. Cases often are selected for presentation *because* they raise special issues, circumstances, and challenges and are one of a kind. If there is any risk that preparation of a research report could reveal the identity of a subject, the subject must be informed of this in advance and provide consent.

Another area in clinical research where invasion of privacy is possible is in the *use of informants* for data collection. Occasionally, treatment research with a client or group of clients may solicit the aid of friends, spouses, neighbors, teachers, or employers. The purpose is to ask these individuals to provide data about the client. The information is used to evaluate the effects of treatment or the severity of the client's problem, although the client may not be aware of this assessment. Seeking information about the client may violate the client's right to privacy and confidentiality.

> The client may not want his or her problem widely advertised, and any attempts at unobtrusive assessment may violate this wish.

For example, asking employers to assess whether the client's alcoholic consumption interferes with work performance may apprise the employer of a problem of which he or she was unaware. The clients or their representative must provide consent in advance before contacting informants unless the informants are legal guardians (e.g., parents of young children; adult children of elderly parents).

The opportunities for invasion of privacy perhaps have never been greater than the present, given the information collected and the interest in using this information for research and other purposes. For example, evaluation of the census every 10 years in the United States provides information that is increasingly sophisticated and accessible. Much of the information is publicly available through databases, which was not as readily accessible earlier. The database will be used by advertisers that wish to target specific neighborhoods, based on income, education, ethnicity, and other variables that can be obtained, sometimes on a street-by-street basis. For example, automobile manufacturers have a profile of the type of persons who buy their cars; advertising can be targeted to neighborhoods where such persons live, as revealed by the census. Many will view this as invasion of privacy or as a slippery slope leading to such invasion, despite the fact that they cannot be identified individually. Researchers ought to be sensitive to these issues and to address risk that might be associated with the seemingly innocuous goal of gathering information.

> Of course much more sophisticated than the census is the automatic tracking of Internet behavior among all of us as individuals.

Without the details revealed to us, our every stroke and Web site visit can be tracked, stored, and directed to advertisers and databases for later use. All of this is done without our explicit consent, although we often click "accept" without knowing that we may have agreed to share that information.

16.4.8: Further Considerations Regarding Special Circumstances

Invasion of privacy often is discussed at the level of the individual subject. However, much larger units are relevant and of deep concern in research. Invasion of privacy of *communities and cultural and ethnic groups* emerges as well. There are two broad issues here. First, even when the identity of the individual is not known, the results can violate the privacy of a large easily identified group. A dramatic example set the tone for current cautions. Decades ago a study was designed to survey alcohol use in an Inupiat community in Barrow, Alaska (Foulks, 1987; Manson, 1989). The purpose was to examine cases of alcohol abuse and to evaluate community detention programs for acute alcohol detoxification. A representative sample of persons (N = 88) over the age of 15 was drawn from the community and interviewed regarding their attitudes, values, and behavior in relation to alcohol use. Other measures of functioning were assessed as well, including church membership and social and work behavior. So far, the project seems innocent enough. However, this all changed with the reporting and dissemination of the results.

The results indicated that 41% of the sample considered themselves to be excessive drinkers; over 50% said that alcohol use caused problems with their spouse and family; 62% said they regularly got into fights when drinking. These and similar types of descriptive statements indicated that alcohol use was a problem in this community. Reports of the findings were viewed by the community as highly objectionable and invasive. The community's view was that alcohol use and associated problems resulted from a new way of life imposed on them rather than on implied deficits, biological or otherwise, or problems inherent to the people. The report was criticized as denigrating, culturally imperialistic, and insensitive to the values of American Indian and Alaskan native culture (Foulks, 1989). Great oversimplification and distortion of the findings by the news media (e.g.,

a byline stating, "Alcohol Plagues Eskimos" in the *New York Times*, January 22, 1980) and emphasis of alcoholism and violence in various articles exacerbated the problem. In relation to invasion of privacy, individual community members could not be identified by the report. Nevertheless, community members, whether or not they served as subjects, viewed their privacy as violated and objected that they were misrepresented (see Manson, 1989). Such examples convey that investigations do not merely describe relations and report findings of abstract scientific interest. The methods of obtaining information, the reporting of that information, and the way information is and could be used are part of the ethical considerations of research.

Second and more broadly pertains to the study of different cultures and ethnic groups and multiple sensitivities that are required. The discussion of this chapter provided generally accepted:

- Definitions of anonymity
- Confidentiality
- Invasion of privacy

But "generally accepted" by whom? The implication is that these terms are neutral, descriptive, and objective in some way. Actually, the terms are very culturally bound.

> Some groups more than others might view the very act of research as an invasion of privacy and object to the questions, topic, and the variables that will be observed or studied.

Thus, one culture might see questions about sexual activities and finances as an invasion; other cultures might well see questions about one's family, past, or beliefs about tradition as an invasion. I mentioned previously that universities and institutions have Institutional Review Boards to review research and that review is designed, among other things, to protect subjects. That may or may not be enough particularly in work with diverse cultures where unwitting insensitivity to privacy issues could readily occur. The case mentioned previously on the study of alcohol use and the Alaska community is one illustration. More protections might be needed where the information can readily identify a group and where the culture needs to define what is and is not appropriate invasion of privacy and group protection.

Diverse cultural groups have learned (the hard way) about the need for protection and have special guidelines and procedures in place. For example, some Native American nations have their own review boards and codes for research (e.g., Ho-Chunk Nation, 2008; Navajo Nation, 2009). The Navajo Nation codes, for example, encompass all research included with their population. The review criteria include more careful protections in some domains than university review boards ordinarily require. Examples include no expedited (streamlined) review of research, the right of the Navajo to negotiate procedures and methods of any project, and of course the right to reject and approve of proposals that will draw on the Navajo nation whether or not the proposal has been approved by some university or other organization (Brugge & Missaghian, 2006). Impetus for such guidelines stems from concerns that researchers often do not understand the problems the tribes are experiencing, are condescending, bring stereotypic views, and do not respect cultural norms.

Culturally sensitive research begins with active participation with the cultural group (Harding et al., 2012; Manson, Garroutte, Goins, & Henderson, 2004; Thomas et al., 2009). Active participation is direct involvement of community leaders at the proposal stage and to address any facet of research (e.g., communicating with and compensating participants, assessment domains and measures), including the consequences that might stem from communication of the findings and use of the information. I have drawn on works with Native Americans to illustrate key issues but of course the rights, protections, and processes apply to any group with its own structure, leadership, community identity, and practices (Ross et al., 2010). That may extend to religious groups as well as cultural and ethnic groups where seemingly innocent or descriptive findings might readily be considered stark invasions of privacy and deleterious in some way.

Many years ago, cross-cultural research was a specialty area within psychology and cocooned in its own journals and professional societies. These journals and societies are alive and well, and cross-cultural studies continue to flourish. Arguably what have changed are increased attention, sensitivity, and recognition of the importance of diversity. Diversity in this context refers to culture, ethnicity, but also to recognition of many groups within a culture, particularly those who are subject to discrimination, harassment, or neglect (e.g., as a function of sexual identity, physical or mental disability). In many cases, this has been mandated by federal law providing individual protections and legal recourse for discrimination and harassment, but there is more here in relation to science. We now recognize that culture and ethnicity can be a moderator (remember—this is variable that can influence the direction or magnitude of the relation between other variables).

> Many core psychological processes (e.g., perception, learning), reactions to assessment, and clinical dysfunction (e.g., dyslexia) can vary as a function of culture.

As more mainstream research attends to culture and ethnicity, more protections of culture and ethnicity too may become more mainstream in regulations that guide research and that are intended to protect subjects.

16.5: Informed Consent

16.5 Evaluate the practice of Informed consent of subjects as used in statistical experiments

We have been discussing cultural groups and that is a special case with its own issues as I have noted. Let us return to the individual participant in research and the requirements here. A pivotal protection is informed consent, namely, that the participant is informed about the project and its procedures and implications and agrees to participate. It is not quite that simple.

16.5.1: Conditions and Elements

An ethical requirement of research is that investigators obtain informed consent before subjects serve in the study. There are occasional exceptions such as situations in which archival records are used and subjects are no longer living or cannot be identified. Implementing the requirement raises special obstacles. In principle, consent can never be completely informed. All possible consequences of the experimental procedures, measures, and participation cannot be known and hence cannot be presented to inform the subject. Also, the impact of the experimental manipulation or intervention, however seemingly innocuous, can have multiple effects (e.g., direct and side effects). Consequently, all the more is it difficult to provide complete information about the intervention and its effects.

Stating the logical status and limits of available information that could be presented to the subject is important as a backdrop for the tasks of the investigator. Information cannot be complete. Yet the responsibility of the investigator is to provide available information and reasonable statements of the likely repercussions from participation so that the subject can make a rational decision. Broad concepts encompassed by informed consent include:

- *Competence*
- *Knowledge*
- *Volition*

These are summarized in Table 16.2 for easy reference.

Competence refers to the ability to understand and engage in decision making about the intervention options.

Characteristics of the sample may impede decision making (e.g., very young or very old subjects, individuals with autism that may impede cognitive functioning) and meeting the competence criterion. The competence of others who act on behalf of the client (e.g., parents, other relatives) is then the issue because these persons take over responsibility for decision making. Obviously, having others make critical decisions can raise its own problems if those who provide consent do not have the interests of the client at heart. (This is another good reason to be very nice to your relatives.) Such circumstances arise in considering invasive or risky medical or psychiatric procedures and do not pertain to the vast majority of interventions we have discussed. Yet, the criterion is the same—does the participant or those who represent the participant really understand what is involved by participating?

Table 16.2: Three Elements of Informed Consent

Elements	Description
Competence	The individual's ability to make a well-reasoned decision and to give consent meaningfully. Are there any characteristics of the subjects or the situation in which they are placed that would interfere with their ability to make thoughtful, deliberative, and informed decision?
Knowledge	Understanding the nature of the experiment, the alternatives available, and the potential risks and benefits. Is there sufficient information provided to the subject, and can the subject process, utilize, and draw on that information? Competence to use this information is relevant as well.
Volition	Agreement to participate on the part of subject that is provided willingly and free from constraint or duress. Are there pressures, constraints, or special contingencies, whether explicit or implicit, that coerce subjects to serve in the study? Penalties or alternatives that are likely to be viewed as aversive for not participating may be a sign that participation is not completely volitional. Also, subjects must be free to revoke their consent at any time.

Knowledge, the second element of consent, pertains to information about the project. To provide adequate knowledge for informed consent, investigators are obligated to describe all facts, risks, and sources of discomfort that might influence a subject's decision to participate willingly. Disclosure of information should be provided in understandable language so that the participant can make an informed decision. All the conceivable risks need not be described, but rather only those that might plausibly result from the procedure. The information must be presented to clients (or their guardians) in an easily understandable fashion. In addition, clients should be allowed to raise questions to clarify all of the issues that might be ambiguous.

Volition means that the subject agrees to participate without coercion. Participation in the experiment cannot be *required* to fulfill a class assignment, according to current requirements for consent. For subjects to provide consent, they must have a choice pertaining to their involvement in the investigation. The choice cannot be one in which participation in the experiment is substituted for some aversive or coercive alternative (e.g., completing two extra term papers), although in any given case, this may be difficult to discern. Whether the subject can "freely" choose to participate is

sometimes evident from the consequences for not agreeing to participate or from withdrawing once consent has been provided. The absence of any penalty partially defines the extent to which the subject's consent was voluntary.

16.5.2: Important Considerations

Competence, knowledge, and volition are not straightforward criteria determining whether consent is informed. Consider a few salient issues. Competence to provide consent is a major concern with populations that may be incapable or less than fully capable of providing consent (e.g., fetuses, young children, persons with intellectual impairment, comatose patients, and institutionalized populations such as prisoners). Determining whether individuals are competent to provide consent presents many problems in its own right, and there is no single, defensible method to do that. That is, there is no standard, agreed-upon "competence" measure that can be administered and scored. Even if there were, the cutoff score might be endlessly debated.

In principle, ensuring competence could be a major issue. In practice, there are large segments of research where this is not an issue. For example, in laboratory studies with college students with psychological tasks (e.g., listening to tapes of innocuous interactions, reading passages and remembering details), surveys of most individuals, and Web-based experiments, competence is not an issue. Subjects are considered quite capable of making rational decisions to participate on the basis of the information provided, and few would be worried about any deleterious effects of participation.

Ensuring that consent is based on knowledge provided to the subject has some ambiguities too. The risks and potential benefits of the intervention are not always well known, particularly for populations that have been refractory to conventional interventions. Last-resort or experimental techniques (e.g., brain surgery to control otherwise-unmanageable seizures, highly experimental drugs for cancer) may be improvised.

There is a way in which consent can never be completely informed. All possible consequences of the experimental procedures, measures, and participation cannot be known for any given individual and hence cannot be presented to inform the participant. This is more easily illustrated outside of psychology where even well-established interventions can have horrible side effects. For example, the oral polio vaccine is used outside of the United States in places where polio is common.[3] The oral vaccine is better suited for widespread distribution and for stopping the spread of polio. Yet, approximately 1 of every 750,000 people who receive the vaccine contract polio from it (Kew, Sutter, de Gourville, Dowdle, & Pallansch, 2005). Complete information is not available to tell individuals whether they are likely to contract polio, so they can use more refined information to evaluate whether they are at high risk. The responsibility of the investigator is to provide available information and reasonable statements of the likely repercussions from participation so that the subject can make a rational decision.

Volition also raises special issues. For example, whether institutionalized populations can truly volunteer for the intervention or research is a potential problem. Individuals may agree to participate because they feel compelled to do so based on real or perceived pressure from others including an investigator.

> Participants may anticipate long-term gains from staff and administration whose opinions may be important for status or commodities within the institution or for release from the institution (e.g., parole from prison).

The lure of release and the involuntarily confined status of many populations for whom the intervention is provided may make voluntary consent impossible.

16.5.3: Additional Important Considerations

Also, monetary inducements can introduce concerns about volition. Many studies pay subjects for participating or for completing assessments and sometimes the amount of money is high (e.g., $200–$500 for completing the assessment battery) for going through various scanning devices and quite high in light of the income of the participants (individuals on welfare or with income at or below the poverty line) (see Dominguez, Jawara, Martino, Sinaii, & Grady, 2012; Grady, 2012). Subject payment usually is framed as reimbursement for time spent in the project, but monetary inducements are clearly a gray area. One can be said to always have a choice of saying yes or no, but one might argue this is a superficial analysis of choice. If some external (or perhaps internal) influence increases the probability of participating to such a high degree, then the likelihood of adopting an alternative (not participating) approaches zero.

Perhaps more subtle than monetary inducements, the differences in power and status in the experimental setting between the investigator or experimenter and subjects can militate against voluntary consent. Subjects may not feel they can choose freely to participate or to withdraw because of their position in relation to the investigator. Since the investigator structures the research situation, the subject depends almost completely in the information provided to make choices about participation or continuation in the investigation. Thus, consent at any point in the research may not be completely informed because the subject may not have access to important information. There are broad group-based issues that raise novel variants about voluntary consent:

- Large databases (e.g., on health, education, genetic data, social behavior on the Web, purchases) are available now like never before.

- The scientific advantages and opportunities also are available like never before either.
- Critical questions can be answered by combining large databases and using them in ways to which subjects did not agree.

For example, in Iceland (but other countries too) careful genealogical data and health records are obtained for the population. Already important findings have been obtained relating DNA to disease, but none of the individuals provided consent for the use of the information in this way. The databases are being combined in ways that none of the participants were told about. The issue of whether consent is needed has placed the matter in the courts and the body that oversees bioethics for the country (Kaiser, 2013a). The most recent rulings deny further mining of the database given that individuals did not voluntarily consent. I mention this example to convey that critical issues of consent are still alive and well and no doubt will continue to be as advances in assessment and utilization of data may raise different opportunities to combine information (e.g., Wolf, Annas, & Elias, 2013). Placing more information on electronic records (health, traffic violations, academic performance, credit card expenditures, Web sites visited, delinquent payments, and more) will make data merging and reporting on segments of the population (including our individual neighborhoods or streets) possible.

Informed consent has become the central issue for ensuring the protection of the individual client. Before participating, information is conveyed about the procedures, likely benefits and possible side effects. As part of the protection, clients are assured that they may stop their participation at any time. Thus, the option of terminating must rest with the client or those who provide consent on the client's behalf. Even when consent can be sought and obtained from the persons themselves, it is often unclear whether consent is adequate or meaningful, particularly for special populations where competence may be questioned.

A key question is whether participants in fact understand the consent forms and procedures to which they have agreed. For example, in some studies of clinical trials, approximately 30–45% of patients do not understand the information that has been provided to them, thought that their treatment was established rather than experimental, and did not know they were assigned to treatment or placebo conditions on a random basis (Flory & Emanuel, 2004). It is important not to take specific percentages I have noted too seriously because understanding what one has signed varies as a function about what facet of the procedure is evaluated (e.g., did the patients understand the goals of the study, the process of randomization, voluntarism, opportunities to withdraw, and the risks and the benefits of treatment) as well as the interventions (e.g., psychotherapy, surgery) and of course the population (e.g., age, type of disability) (e.g., Falagas, Korbila, Giannopoulou, Kondilis, & Peppas, 2009). The quandary is that informed consent usually refers to *procedures* to which subjects are exposed (e.g., a speech and consent forms) rather than *an outcome of the procedures* (e.g., whether subjects in fact know, understand, and fully appreciate all the information the consent procedures were designed to convey). As a procedure, informed consent is fairly easy to obtain; as an outcome, i.e., that people really understand exactly what they are getting into, is another matter.

Many efforts have been made to improve communication to patients about the procedures and other facets of the study in light of the data on very limited comprehension about the study, procedures, and risks (e.g., Schenker, Fernandez, Sudore, & Schillinger, 2011). But it is not a matter of communication alone. Many consent forms omit key features (e.g., full statement of risk, rights to withdraw, potential conflict of interests of investigators) that are supposed to be presented (Palmour et al., 2011). Also, many of the techniques used in clinical research (e.g., Positron Emission Tomography [PET] and Single-photon Emission Computed Tomography [SPECT]) are more difficult to explain than some of the more modest methods of the past (e.g., filling out the Beck Depression Inventory).

16.5.4: Consent and Assent

Informed consent is the process and procedure that permits participation in research, and our discussion to this point has assumed that adults were the participants in the study. Research conducted with children and adolescents raises additional issues. In this latter research, informed consent usually is provided by a parent or guardian. Yet, if the children are old enough to understand the proposed research and activities expected of them and perhaps risks and benefits, "assent" is sought. *Assent consists of being willing to participate in research.*

> For the assent criterion to be met, the child must affirmatively agree to be involved in the research project.

The affirmative agreement part is pivotal, and merely not objecting to being in the study does not count.

In the United States, Federal regulations define children as persons who have not attained legal age for consent (www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.408). In most states in the United States, this is under the age of 21 (but 18 in some states).

> These regulations require that assent be obtained directly from the child or adolescent unless the child is incapable because of immaturity or cognitive inability to understand the procedures.

Formal documentation is required. Thus, there is an "assent form," for children, just as there is an "informed consent form" for their parents, as illustrated in the next section.

Assent is in addition to informed consent of a guardian and not a replacement for that consent. In any given situation, child assent may be waived (e.g., if there is an urgent health need to be met of the child that is only available in a research project) and local resources (e.g., university review committees) are allowed discretion in deciding whether assent is feasible in special circumstances. Yet here too the default position is obtaining assent from a child. Committees that oversee and approve of research are responsible for making the decision or approving the investigator's request if the case is made not to seek assent.

16.5.5: Forms and Procedures

In advance of placing subjects through any procedures or assessments, a consent form is provided to convey information about the study that the subject ought to know to make an informed decision. Usually, Institutional Review Boards and committees (e.g., at colleges, universities, hospitals, prisons) are charged with evaluating the research proposal, consent procedures, and consent form. Members who review the proposal are drawn from diverse disciplines. The research proposal is evaluated to examine the research design, specific procedures, the conditions to which the subject will be exposed, and risks and benefits. Evaluation of the research design deserves comment. The general plan of the research must be made clear to permit committee members to determine if the questions underlying the investigation are reasonable and can be answered by the study. If the questions cannot be answered by the study, then the subjects should not be placed at any risk or inconvenience. Methodological scrutiny is not very stringent, but it need not be at this point. The investigator ought to be able to make the case that the study is worth doing and that the ends (the results) justify the means (procedures to which subjects will be exposed). In some cases of course, the means will not be allowed no matter what ends, in other cases the ends seem trivial and hence any means (or use of subjects) may not be worthwhile. (In my dissertation proposal, no one could tell if there were ends or means, so my proposal breezed through the consent committee.)

Most psychological experiments (e.g., with college students in laboratory studies of psychological processes such as cognitions, memory, attributions) do not involve risk situations and are designated as "minimal" risk. The subjects (e.g., college students), experimental tasks (e.g., engaging in a cognitive task on a computer touch screen, completing personality measures), and risks (e.g., mild boredom if that task continues too long) do not exceed the risks of normal living. Review of such studies is relatively straightforward because concerns about subject rights are not raised by the research paradigm. In many such cases, formal review of the study and informed consent procedures are omitted because the experiment is considered to be in a class of procedures that is innocuous. Essentially, such procedures are given blanket approval. More likely they are given a quick review (expedited) in light of core features such as minimal or no risk and subject anonymity.

In clinical work, several features often extend the situation well beyond "minimal risk" by virtue of the population (e.g., patient samples), focus of assessment or intervention (e.g., suicidal intent, depression), and special ethical dilemmas (e.g., random assignment, delaying treatment), as discussed further below. Understandably, the review of proposals and consent procedures of such studies are more stringent. In many universities, separate review committees are available for different types of research. For example, a social sciences review committee often reviews psychological experiments with minimal risk. In contrast, research with clinical populations may be more likely to be reviewed by a biomedical committee.

Providing consent is operationalized by the subject's being told about the study and then completing the consent form. Federal regulations specify eight elements that comprise informed consent, and these are presented in Table 16.3. The elements are elaborated more concretely in

Table 16.3: Eight Basic Required Elements of Informed Consent Materials Presented to Research Participants

Basic Required Elements of Informed Consent Materials
A statement that the study involves research, an explanation of the purposes of the research and the expected duration of the subject's participation, a description of the procedures to be followed, and identification of any procedures that are experimental.
A description of any reasonably foreseeable risks or discomforts to the subject.
A description of any benefits to the subject or to others that may reasonably be expected from the research.
A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject.
A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained.
For research involving more than minimal risk, an explanation as to whether any compensation and an explanation as to whether any medical treatments are available if injury occurs and, if so, what they consist of, or where further information may be obtained.
An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights, and whom to contact in the event of a research-related injury to the subject.
A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits to which the subject is otherwise entitled.

Source: U.S. Department of Health and Human Services (2009; www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.116).

Table 16.4: Components of Informed Consent Forms

Section of the Form	Purpose and Contents
Study Title, Principal Investigator, Funding Source	Basic information about the study, who is in charge, and source of funds if any
Solicitation of Participation	Stating that the person is "invited" to participate in a research study
Description of Procedures	Presentation of the goals of the study, why this is conducted. Clarification of the experimental conditions, assessment procedures, requirements of the subjects
Risks and Inconveniences	Statement of any physical and psychological risks and an estimate of their likelihood. Inconveniences and demands to be placed on the subjects (e.g., how many sessions, meetings, requests of and contacts with the subjects)
Benefits	A statement of what the subjects can reasonably hope to gain from participation, including psychological, physical, and monetary benefits
Costs and Economic Considerations	Charges to the subjects (e.g., in treatment), payment (e.g., for participation or completing various forms); any compensation to the subject
Alternative Treatments	In an intervention study, alternatives available to the client before or during participation are outlined; that is, if the client does not agree to participate to receive the conditions of this study, other options are available
Confidentiality	Assurances that the information is confidential and will only be seen by person(s) who need to do so for the purposes of research (e.g., scoring and data analyses), procedures to assure confidentiality (e.g., removal of names from forms, storage of data). Also, caveats are included here if it is possible that sensitive information (e.g., psychiatric information, criminal activity) can be subpoenaed
Selective Refusal	A statement noting that the subject does not have to answer any particular question or complete a measure if he or she does not want to
Voluntary Participation	A statement that the subject is willing to participate and can say no now or later without penalty of any kind
Questions and Further Information	A statement that the subject is encouraged to ask questions at any time and that the answers received were satisfactory
Contact Person	A statement that the investigator can contact a person or persons (listed by name and phone number) who are responsible for the project
Signature Lines	A place is required for the subject to sign as well as for the experimenter
Authorization and Approval	A stamp of approval or equivalent that the consent form has been approved and dated by the institution overseeing research

NOTE: Many institutions (e.g., universities, hospitals) have sample consent forms. Also, many of these can be obtained on the Web via a search engine and typing in "informed consent form for psychological research" or for "clinical trials."

the Table 16.4, which lists the likely sections that will be included in a consent form and serves as a translation from the federal regulations to move closer to what is provided for the subject. From these sections, it is fairly easy to devise a specific consent form. The material in the form is explained to the subject verbally and the subject then can read and sign the form indicating that she agrees to participate.

I mentioned that in the case of research with children, assent usually is sought. The child has the opportunity to agree to be in the study or to withdraw. The structure of the assent form presented to the child closely follows the structure of the informed consent form presented to the parent. The key difference is in providing the key content in a language that is more understandable. Table 16.5 provides a sample of some of the sections and wording that is likely to

Table 16.5: Selected Sections of an Assent Form to Illustrate Wording for Children

Section	Description
Invitation	We are asking you to be in a research study. This form will tell you all about the study and help you decide to be or not to be in the study. Read this paper carefully and ask any questions you have. You might have questions about what you will do, how long it will take, if anyone will find out how you did. When we have answered all of your questions, you can decide to be or not to be in the study. This is called "informed consent."
Confidentiality	If you participate in this study, we will not tell anyone else how you did. We will keep all information about your participation in a locked cabinet without your name on it so that only we can see how you did. We will use this information to write a big paper about the study. Your name will not be used in that paper. After we write the paper, we will throw away all of this information.
Your Rights	You have the right to choose not to be in the study, and nobody will be mad at you. You have the right to stop participating anytime you want, and you will still get the prize.
Consent	Signing this paper means that you have read this or had it read to you and that you want to be in the study. If you don't want to be in the study, don't sign the paper. Remember, being in the study is up to you, and no one will be mad if you don't sign this paper or even if you change your mind later.

NOTE: These sections have a template that is made available on the Web by the University of Maryland Baltimore County (see General instructions and sample to create an assent form at www.umbc.edu/irb/sampleassent.doc. Material in brackets in the form note special places where changes or modifications may be needed).

be at the level used in an assent form. Of course age of the child and any limitations in understanding would influence the exact wording and indeed whether or not assent would be used.

> Both informed consent and assent forms are submitted for formal review by an Institutional Review Board in the setting (e.g., university).

That board is an appropriate resource to discuss assent and whether it might be waived under special circumstances.

16.5.6: Certificate of Confidentiality

In clinical research, sensitive information may be collected and greater protection can be provided to the participants beyond those specified by informed consent forms. Specifically, a Certificate of Confidentiality can be provided. What this is and the goals are clearly stated as follows:

"**Certificates of Confidentiality** are issued by the National Institutes of Health (NIH) to protect identifiable research information from forced disclosure. They allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level. By protecting researchers and institutions from being compelled to disclose information that would identify research subjects, Certificates of Confidentiality help achieve the research objectives and promote participation in studies by helping assure confidentiality and privacy to participants" (http://grants.nih.gov/grants/policy/coc/, quote from the Web site, NIH, 2013a).

Use of the Certificate is likely in situations where sensitive information is disclosed as part of the study. The type of information that is sensitive might relate to:

- Sexual attitudes, preferences, or practices
- Information relating to the use of alcohol, drugs, or other addictive products
- Information pertaining to illegal conduct
- Information that, if released, might be damaging to an individual's financial standing, employability, or reputation within the community or might lead to social stigmatization or discrimination
- Information pertaining to an individual's psychological well-being or mental health
- Genetic information or tissue samples (http://grants.nih.gov/grants/policy/coc/faqs.htm#365)

Clearly, participants could be reluctant to participate for fear that such information might be obtained by others. The Certificate is designed to allay or reduce that reluctance. Most research in clinical psychology and related fields does not utilize a Certificate of Confidentiality. This is an extra step for protecting subjects.

> Any study in which sensitive information is collected and the subjects' identity is collected as well in principle is eligible for this Certificate.

As a passing comment, my own work relies on children referred for clinical problems (children referred for conduct or oppositional defiant disorder) and it is often the case that parents wish to protect critical information given questions we ask (e.g., past prison experiences, domestic violence, use of drugs, current conflict and abuse with their children). In light of the scope and nature of these questions, the Certificate provides reassurance to participants (but also to me as an investigator) that I will not be forced to disclose information against everyone's will.

As with informed consent, participants are fully informed about the protections provided by the Certificate. This is particularly important because the protections have exceptions. For example, if there is evidence of child abuse or danger on the part of the participant, the researcher would be required to disclose that. Information on protections and exceptions is part of the procedure of presenting and signing the Certificate of Confidentiality. Overall, the Certificate is a valuable protective step for participants.

> Is it absolutely foolproof and an ironclad guarantee for protection?
>
> Very few protections would meet that bar all of the time. On occasion squabbles with the court and back-and-forth rulings have broken the confidentiality (e.g., Beskow, Dame, & Costello, 2008). This is clearly an exception but is important to mention.

16.5.7: Letter and Spirit of Consent

Concretely, the investigator is required to describe the procedures to the subject and to obtain signed consent. The signed consent form satisfies the research requirements and hence follows the "letter" of the rules that govern research and the investigator's responsibilities. In addition, there is a "spirit" of informed consent, which refers more nebulously to the overall intent of the procedures and the goal to ensure that clients genuinely understand what they are signing, what the study entails, and the risks, costs, and benefits. In most research (e.g., laboratory studies with college students), presentation of the consent information followed by the subject's signing of the form is sufficient for the letter and spirit of the consent procedures. Perhaps in that context, if there is any conflict of letter and spirit, perhaps it arises in conveying that the subject is free to withdraw without penalty. That is true (letter) but the alternative assignment that the students need to complete to satisfy the course (introduction to psychology) requirement can vary (e.g., a small term paper, reading and reporting on other experiments) and

to the student if not to external observers that requirement might be argued as coercive. These matters do not concern too many people it seems because on a continuum of coercion in experiments, this may not register. So if the letter is met here, few want to debate the spirit of noncoercion about alternative assignments. However, ethical concerns have been voiced about the practice of using and recruiting introductory students (Leentjens & Levenson, 2013). Moreover, a significant portion of students feel coerced by the recruitment procedures (e.g., Miller & Kreiner, 2008).

Research that is any way service related (e.g., treatment, rehabilitation, special visits, or care) or that involves personally or physically invasive procedures (e.g., obtaining private information that could be solicited by the courts, medical tests with risks) or participants who are or may not be competent to represent themselves fully (e.g., children, disadvantaged persons, individuals with intellectual disability or psychiatric symptoms) raises special obstacles. Clients may be less likely to understand options, choices, and opportunities to change their minds about participation. Indeed, I mentioned previously how little understanding select groups may be of experimental procedures and the conditions of experimentation. In such cases, satisfying the letter of the informed consent requirements may not approach the spirit or intent of these requirements.

Interestingly, there are no formal research requirements that subjects actually understand what is presented to them. Research so often has shown that clients do not understand key features of consent that a cynical view is not a great leap. That view is that emphasis of informed consent is on getting the signature on the form and for the protection of the institution and researcher (against litigation) as much if not more than the participant. Perhaps that view is too extreme. I believe a more data-supported view would be to note that presentation of information and signing of consent forms might be accomplished without genuinely informed consent. For this reason, both the spirit and the letter of consent are important.

> The spirit of consent emphasizes the investigator's responsibility to maximize the clients' understanding of the investigation and refers to investigator's "best effort" to convey the purpose, procedures, and risks of participation and generally to meet the consent conditions.

It is not a disaster if a prospective subject occasionally refuses to participate. In fact, this may be a good index that the procedures to explain the project are registering. This is not our usual way of thinking about consent in which a subject who refuses may be viewed as a recalcitrant, odd, vegetarian who eats a lot of beef.

Presentation of the content by repeating significant facets of the study, paraphrasing the consent form, asking questions of the subject at critical points to assess understanding, and similar strategies may help foster better comprehension. The time allowed to explain the procedures and to obtain consent may need to be extended to foster the atmosphere required to inform the client. Obviously, protracted consent procedures, clinically *over* sensitive presentations, and ad nauseam requests for feedback on the part of the experimenter. So it is safe if not wise to stay away from such phrasing as, "You are probably wondering what 'random' assignment really is, why we do this, and whether assignment can ever be truly random. Did you ever hear of R.A. Fisher—I didn't either . . . .", or "O.K., how do you feel about (or 'are you comfortable with') pressing on the touch screen to select between these pairs of words that will be presented? No one has sprained a finger before when pressing the screen but it is possible. I'll bet this is a little worrisome." These comments to the participants move too far but could make the consent procedures more interesting (and educational) than the experiment itself. The rules or letter of consent procedures specify the minimal conditions to be satisfied in discharging responsibilities to the subject. Beyond that, judgment, experience, and common sense are needed to meet that the goals of consent and to balance research interests and subject rights.

The spirit of consent is important to underscore for a reason not frequently acknowledged. The investigator often has a conflict of interest in obtaining informed consent. Entry of subjects into the study is the goal and something critical to the investigator (e.g., meeting the demands of a grant, completing a study for a thesis or dissertation, a possible publication that may result) all fall on the side of getting a subject to say yes and participating without hesitation. That pressure to enter subjects into the study is readily passed on to research staff or assistants who are the ones actually obtaining consent. On the other side, there is an obligation to present the information and meet the letter and hopefully the spirit of consent. As consent moves from perfunctory to detailed, more time is required with the subject and more risk or perceived risk on the part of the investigator of losing the subject increases. This is a natural tension of the investigator to be aware of and to combat to provide thoughtful explanations of the study.

16.6: Intervention Research Issues

16.6 Investigate ethical issues in applying Intervention practices on subjects of statistical studies

In studies of various interventions such as psychotherapy, counseling, and education, additional ethical issues arise

or nuances emerge that warrant consideration. Several issues emerge that may vary with the type of:

- Intervention (e.g., treatment vs. prevention)
- The population (e.g., young children, hospitalized adults)
- Setting (e.g., university, patient services)

16.6.1: Informing Clients about Treatment

An important issue is the information that is provided to the client about the intervention. Outside of the rationale and procedures themselves, the investigator is required to convey the current status of the treatment, assuming that the client is able to understand the information. Whether treatment has been shown to be effective or not in previous applications would seem to be important and routine information. Many treatments are experimental, and the subject normally can be provided with a statement to that effect.

Therapy research raises an interesting dilemma because honesty about the basis of treatment might attenuate some of the therapeutic effects. The processes through which therapy achieves its therapeutic are not really known. However, mobilization of hope and expectancies for change in the client are among the mechanisms proposed to contribute to, if not largely account for, change (e.g., Duncan, Miller, Wampold, & Hubble, 2010; Frank & Frank, 1991; Lambert & Ogles, 2013). These factors are common to many treatments and hence could explain why many different treatments work and why people get better with attention placebo conditions and placebo medications. Mentioning the current status of treatment and the possibly important role of hope and belief in the procedures might well attenuate the impact of these factors.

What does one say to the prospective client/subject? "Oh yeah, we do not know that this treatment works or how it works if it does, but hope could be a big factor. That is, it is what you believe as much or more important than what we do and could make the big difference? In fact, some experts believe what we do is not too important as long as you are convinced it is something."

Suspicions about treatment efficacy might be raised by full disclosure. In some treatment studies, the independent variable is the expectancy for success conveyed to the subjects. Hence a treatment is claimed to be very effective or ineffective, depending upon the condition to which the subject is assigned. Disclosure of the current status of the technique would compete with this manipulation.

Information about treatment in an experiment may extend to the treatments the subject will not receive. Conceivably, subjects could be told that there are different treatments, only one of which they will receive. Subjects might want to know whether some treatments are more effective than others and whether they have been assigned to a "control" group. In addition, subjects may show a clear preference for an alternative treatment and react adversely to the condition to which they are assigned. As alternative treatments become known, skepticism about a particular treatment may arise and therapeutic improvement may be affected.

At the beginning of the study, subjects ought to be told that there are various treatments offered and that assignment to treatment is random, assuming that these are in fact the case. Although subjects are rarely pleased to learn that their assignment to treatment will be random, the importance of randomness in assessing the impact of different treatments might be stressed. Only those subjects who agree to the conditions of the investigation can serve as subjects and be assigned to conditions. This does lead to selection of a special group and that may have implications for external validity of the results. Moreover, even among those who agree, they dropout quite selectively based on the condition to which they are assigned. Understandably, those assigned to a control condition (e.g., wait list, routine care) can discern that may be more likely to dropout in higher numbers than those assigned to the experimental condition or treatment. In many trials (e.g., medication, surgery) the active and control conditions are not discernible by subjects (or staff) (e.g., placebos are packaged and look exactly like the medication; sham surgery with incisions, anesthesia, and recovery similar to the "real" operation). In trials of psychotherapy placebo control conditions are used less often than treatment as usual controls to control for participation in treatment, to provide ethically defensible care (what individuals usually receive), and to minimize dropping out of treatment that a less credible or fake intervention might promote.

16.6.2: Withholding the Intervention

Intervention studies often withhold the special treatment or preventive intervention and assign some of the subjects to no-treatment or waiting-list control conditions. Although these control conditions are essential to answer specific research questions, their use raises obvious ethical questions. Assigning a client to one of these conditions withholds treatment from which a person may benefit. At the very least, treatment for the client is delayed. If the client's condition does not deteriorate, the delay has increased the duration of misery that may have precipitated seeking treatment. At the worst, the client's condition may deteriorate during the period when treatment is withheld.

The poster child study for flagrant ethical violation based on withholding treatment is the study in the United States (Tuskegee, Alabama) from the 1930s to 1970s. The Tuskegee Syphilis Study, as it is known, was conducted by the U.S. Public Health Service, evaluated the long-term effects of syphilis (see Gray, 1998). Syphilis is a sexually

transmitted bacterial disease. The infection usually goes unnoticed because the symptoms do not emerge immediately. There are stages of the disease that move from rashes and lesions, through a latency or seemingly dormant period, to later stages of blindness confusion, paralysis, dementia, and eventually death, all taking possibly decades (e.g., 10–30 years) to unfold. The later stages were once interpreted as a mental illness referred to as general paresis. Many psychiatric symptoms emerge including primarily dementia but also depression, delusions, confusion, mania, apathy, mania, irritability, and others. Identification of the bacteria responsible for this explained the symptoms as part of a biological disorder in its late stage.

During the 40-year study, 399 African American men were denied effective treatment of syphilis so that researchers could study the progression of the infection. During and after World War II, especially in the mid-1940s, penicillin an effective treatment became widely available but was still withheld to permit the study of the natural course of the disease. The participants were regularly followed and examined but not given treatment. By the end of the study, 74 participants were still alive; approximately 100 had died from syphilis. In the early 1970s, a journalist reported the story nationally that had far reaching consequences (e.g., Senate hearings, lawsuits filed by families, and payouts to families), including the development of federal regulatory guidelines for the research, law covering legal the rights of subjects to receive recourse, development of review boards to consider research before it is conducted, and perhaps even more salient, attention to racism embedded in the study (Brandt, 1978).

The study and its consequences have lingered way beyond the 1970s in light of the shocking ethical breaches. For example, in the late 1990s, a White House ceremony and presidential apology was provided to survivors and family members. Also, a more contemporary source of fallout has been sustained among many African Americans in relation to the medical community. These suspicions, obviously well placed, have led to reluctance to participate in HIV/AIDs intervention programs and medical research more generally (e.g., Poythress, Epstein, Stiles, & Edens, 2011; Thomas & Quinn, 1991).[4] This example applies to multiple ethical concerns even though withholding an effective treatment is enough. There was a huge deception of the goals, procedures, and options, and this was ongoing over the full course of the study.

In psychological studies, ethical issues lingering from past abuses in medical studies are raised by withholding treatment. An investigator is obligated to consider seriously whether a control condition that delays or completely withholds treatment is necessary for the questions addressed in the research. Because of the ethical problems, it may be more appropriate to reserve questions comparing treatment with no treatment to situations where subjects are willing to wait and are unlikely to suffer deleterious consequences. Obviously, volunteer clients solicited from the community may be more appropriate for a study in which a waiting-list group is required than clients who seek treatment at a crisis intervention center. When clients have severe problems and warrant or demand immediate intervention, questions comparing treatment with no treatment are more difficult to justify and to implement.

In some cases, assigning subjects to a waiting-list control group will not really delay treatment. Waiting lists are common at many clinics. A delay before entering treatment may average a few or several months before clients are seen. All subjects who are to serve in the study and who agree to participate can be moved up on the list. Those who are randomly assigned to the intervention condition are treated immediately; those who are assigned to wait can be assessed and then wait the usual delay period of the clinic before receiving treatment. Ethical issues are not eliminated by rearranging one's status on the waiting list. Moving some clients up on the list may delay the treatment of others who are not in the study.

> Some of the problems of delaying treatment can be alleviated by informing clients at intake of the possibility that they will not be assigned to treatment for a particular (specified) interval.

As noted before, the investigation would only use subjects who agree with this stipulation and then randomly assign them to the various treatment and control conditions.

Interpretation of a study that compares treatment to no-treatment alone can be difficult. We know that merely participating in any activity that resembles something therapeutic even if fake is likely to lead to improvement. Thus, comparing a treatment with no-treatment controls is not very informative. The treatment would be better because of some special feature, but a more parsimonious interpretation is that client expectancies for improvement and contact with a therapist led to change. This limitation has led to greater use of other treatment conditions or treatment as usual in a study, even if a no-treatment or wait-list control condition also is included.

16.6.3: Control Groups and Treatments of Questionable Efficacy

In outcome research, some treatments in a given study might be expected to be less effective than others. The use of treatments that have a low probability of being effective raises an ethical issue for the investigator. The issue can emerge in using groups that are designed to control for common treatment factors, such as attending treatment sessions, meeting with a therapist, and believing that treatment may produce change. These groups are designed with the expressed idea that there are few if any components that will actively help the client. Providing a

treatment designed to be weak or a control condition designed to be ineffective raises obvious ethical problems:

1. The client's problem may not improve or may even become worse without an effective treatment. To withhold a treatment expected to be relatively effective renders these possibilities more salient.
2. Clients may lose credulity in the process of psychological treatment in general. Clients expect to receive an effective treatment and to achieve change. If treatment is not reasonable in their judgment and does not produce change, clients may be generally discouraged from seeking help in the future.

In general, the control conditions ethical evaluation by the investigator and review boards. This of course applies to any special control condition in which the likelihood of improvement is unexpected or minimal. Other contextual issues such as who the clients are (e.g., patients seeking treatment, community volunteers) and provisions after the study is completed (e.g., free treatment and care) may affect evaluation of the issues.

At the beginning of the chapter, I mentioned how concerns over placebo controls emerged in the context of research on the treatment of HIV. The concerns have emerged much more broadly and deserve further comment. For example, in the development of medications for depression, there is a controversy about whether placebo controls should be used. The Food and Drug Administration in the United States requires a placebo control to identify whether a drug improves upon the often potent effects of placebos in a given sample. Placebos can significantly improve depressive symptoms in 30–50% of a clinical sample. One has to show that a medication improves upon this percent. Many medications do, but the increment is surprisingly small (e.g., 10–30%).

Can the placebo procedure be justified? Are patients assigned to this condition at special risk for other problems?

What do you think?

In partial defense of this policy, meta-analyses of several studies have shown that the risk for suicide does not differ among subjects who received to treatment or to placebo control conditions. That has not allayed all the concerns. Perhaps the quality of life of the subjects in the placebo group is not as good as those in the treatment group because of the likelihood of continued symptoms and also because of the stress and anxiety of serving in a placebo condition or in a study with such a condition.

What is an appropriate control condition in intervention research and when and whether to use placebo control conditions are matters of worldwide interest, discussion, and debate. In considering the issues, the World Medical Association (WMA) passed a resolution to revise the 1964 Declaration of Helsinki, based on an initial conference convened in Finland.[5] The Declaration addresses research ethics that emerged in response to gruesome medical experiments during the Nazi era. The Declaration provides ethical principles that are designed to govern experimentation with humans (WMA, 2013). The Declaration is not binding, and several countries and agencies and professions within those countries have their own standards and codes. Yet, the Declaration is a significant statement and designed in part to influence other codes and research broadly on a worldwide scale. The Declaration has been amended several times to expand requirements for care of patients and decisions about the conditions to which they ought to be exposed. The original version did not preclude use of placebos but was not explicit. The most recently revised version states explicitly that placebos may only be used when there are no other therapies available for comparison. If there is a standard treatment (e.g., an approved medication on the market), any new treatment should be compared to that rather than to a placebo.

> Development of codes from a world consortium has not been provided specifically for psychosocial interventions that span the relevant disciplines (clinical and counseling psychology, psychiatry, nursing, and social work) where control conditions are used.

Direct translation of the policy from the Declaration of Helsinki is not straightforward because many of the treatments that are considered standard (e.g., much of psychotherapy) are standard only because they are used a lot and have history and tradition behind them, not because they have evidence in their behalf. With the emergence of evidence-based therapies, perhaps it will be reasonable to extend the stance on placebos to psychotherapy trials. Placebo control conditions might not be justified if there is an evidence-based treatment that could be used for the clinical problem under investigation. Perhaps a plausible even though not evidence-based treatment would be a suitable argument to forego placebo controls as well.

16.6.4: Consent and the Interface with Threats to Validity

Some of the ethical issues in this chapter connect directly with drawing inferences and experimental validity of the experiment. Informed consent procedures illustrate this very nicely.

> Two issues that have broad implications for drawing inferences from research are randomization and attrition.

Mentioned on a few occasions already is the importance of random assignment for all sorts of reasons, but primarily to make implausible a host of threats to internal validity (primarily selection biases). Mentioned also was the fact that in any research in which the subject must come back for more than one session (e.g., as in treatment and prevention studies, longitudinal studies of development), subjects may dropout. Dropping out (attrition) can influence all types of experimental validity.

Informed consent raises issues that affect dropping out and threats to validity. Assume that a study compares treatment versus a no-treatment or an attention-placebo control condition. Subjects are informed that they will be assigned to one condition or the other and that the assignment is random. Subjects want to be in the treatment group. After subjects are assigned, they now may evaluate their status and draw inferences about the group they are in. It is likely that more often than not, they will guess correctly, especially if they have any opportunity to chat with other subjects.

> Subjects might wisely sign the consent form with knowledge that they can withdraw later if they do not get the condition they wish.

How ought the investigator proceed? One option is to provide consent information, to underscore the importance of participation even if assigned to the no-treatment (waiting list) or attention-placebo group, and perhaps conveying that if assignment to no treatment may lead to dropping out, it is better to do so now (before being assigned), although of course the subject may drop out at any time. After all cases agree, then assignment can be made randomly to conditions. This may reduce attrition because dropping out before even being assigned was encouraged. The dilemma is that external validity of the results, i.e., the extent to which the results extend to patients will be challenged further. That is, those who participate may be more restricted in their characteristics because some effort was made to use only those who really said they would be likely to remain in treatment.

The notion that subjects can change their minds at any time and withdraw consent is an important protection for subjects. As they gain more information (knowledge as a condition for consent), they may decide the intervention is not for them. Dropping out affects experimental validity, and this is something to be aware of. Analyses of dropouts, intent-to-treat analyses of the data, and sophisticated mathematical methods of estimating (imputation) the missing data are all designed to address the problem of dropping out of treatment; none is as good (in relation to threats to validity) as retaining participants in the project. That is the rationale for inducements (e.g., monetary, lottery for a special prize) for completion of the study.

16.6.5: General Comments

The ethical issues raised in intervention research depend upon the precise research question and the control or comparison groups that form the basis of the design. Use of no-treatment or waiting-list control groups is essential in research that asks the basic question, "Does this treatment work?" The question usually requires assessing the extent of change without treatment.

> Similarly, use of a nonspecific treatment control group is important in research that asks the question, "Why does this treatment work?"

Such research may require a group to look at the influence of nonspecific or common treatment factors.

The research questions that require ethically sensitive control conditions are fundamental to progress in understanding treatment. The questions themselves cannot be abandoned. However, the conditions under which these questions are examined can be varied to attenuate the objections that normally arise. For example, questions requiring control conditions that withhold treatment or provide nonspecific treatment control groups need not be conducted in settings where clients are in need of treatment and have sought a treatment to ameliorate an immediately felt problem. On the other hand, for situations in which volunteer subjects are solicited and can be informed about the experimental nature of all treatment procedures, a wider range of experimental conditions is more readily justified. In short, when the setting has patient care and service delivery as the higher priority, the use of groups that withhold treatment or present "nonspecific" treatments that are expected to produce minimal change is generally unacceptable. When research, rather than service delivery, has the higher priority and clients can be informed of the implications of this priority, the use of such groups may be more readily justified.

Some of the ethical issues of treatment can be ameliorated by providing all subjects with the more (or most) effective treatment in the project after they have completed the treatment or control condition to which they were assigned. After treatment, clients who served as a no-treatment control group also should receive the benefits of treatment. Indeed, this is exactly what the waiting-list control group receives. In studies with many different treatments or a nonspecific-control condition, clients who are not completely satisfied with their progress eventually might be given the most effective treatment. Thus, clients may benefit from the project in which they served by receiving the better (or best) treatment. From an experimental standpoint, this strategy is useful in further examining the extent of change in clients who continue in the superior treatment. Essentially, there is a partial replication of treatment effects in the design. From an ethical standpoint, providing all subjects with the most effective intervention may attenuate objections against assigning subjects to treatments varying in anticipated effectiveness. Of course, at some point in the research long-term follow-up studies are needed in which we see whether the seemingly effective intervention is better than no treatment long-term. This might be done in a randomized controlled trial or creative use of cohort designs in which groups that have not received treatment are followed.

There are pertinent dilemmas of using various control conditions (e.g., the increased use of treatment as usual as a comparison/control condition). While this has its own dilemmas (including there is "no" apparent treatment as usual in the world), the advantage is that nothing is withheld from clients that they would not usually receive and control for internal validity threats and some construct validity threats (e.g., attention and contact with a therapist, expectancies) are addressed as well. The overall issue is instructive in conveying how the design (e.g., randomization to groups of a true experiment), control of threats to validity (e.g., internal, construct), and how that translates to the groups that are used (no treatment, attention placebo) are embedded with ethical issues and considerations.

16.7: Regulations, Ethical Guidelines, and Protection of Client Rights

16.7 Express the position that the law takes in guiding ethical statistical research

We have taken up several specific issues to convey the ethical issues and responsibilities of investigators. These issues and responsibilities are codified in many guidelines and in the U.S. Federal regulations that elaborate the rules, standards, and responsibilities of researchers. Formal guidelines are needed for at least four reasons:

1. One cannot leave standards up to individual investigators. Individual judgment for decision making can be idiosyncratic and hugely biased even when intentions are great. For example, one may see one's own research as having critically important implications and justify procedures (e.g., deceiving others, withholding treatment) as a reasonable price to pay for the information. Guidelines codify accepted and agreed-upon standards and responsibilities and bring consistency among practices.
2. There is often an enormous power differential in both research and intervention settings in relation to the person in charge of the research or intervention and the person who participates in that study or receives the intervention. The investigator or person in charge of the program has more control over the situation and more information about what can be expected, including risks, sources of discomfort, and any side effects. In some cases, clients are at a disadvantage because of their condition (e.g., clinical dysfunction, age [children, elderly], desperation because survival is threatened or because other interventions have not been effective). Differences in power or position do not necessarily lead to abuse of that power. Yet, it is valuable and probably essential to have some checks to ensure the less-powerful party has some protections. Guidelines can provide those checks and serve as a point of reference for investigators, participants, and third parties (e.g., review committees, lawyers).
3. Many decisions in research and services reflect genuine dilemmas, trade-offs, and weighing of risks and benefits. When to provide a more risky or aversive intervention (e.g., to eliminate self-injury that has not responded to less risky treatments) and to offer limited resources (e.g., to decide who shall receive organs for transplants given enormous waiting lists). No one person usually has the authority, wisdom, or balanced perspective to resolve the dilemma or to make decisions. That is why in one way or another often many people are involved in the decision, including consumers (e.g., potential participants) and professionals with specialized knowledge or training (e.g., researchers, ethicists, lawyers) to craft codes. Guidelines represent the input of many who have deliberated about the ethical but also practical issues that ought to be considered in decision making.
4. Guidelines can readily be revised and updated to handle novel issues that emerge. Science is a moving target in both the topics that are studied and the procedures used to study them. For example, assessment techniques (e.g., neuroimaging), new ways of invading privacy (e.g., one's genetic code), use of materials (e.g., stem cells), and new kinds of therapy (e.g., genetic therapy, nano therapy for cancer) raise new dilemmas, challenges, and often unknown or incompletely known risks. Guidelines need to be reexamined periodically to keep up with novel situations and twists on research findings of what to worry or not worry about.

Ethical guidelines along with several specific practices (e.g., informed consent from the client or those in their charge) are needed to help with decision making and to protect the interests of the client or participant.

> Various guidelines, rules, laws, and policies are designed to protect against abuses or lapses in guaranteeing human rights as well as setting high positive standards of what constitutes quality care.
>
> The guidelines are efforts to codify accumulated wisdom from experience and combined with principles of how one ought to act in professional contexts. The guidelines provide explicit standards to which professionals can be held accountable. They are also designed to foster transparency of what is being done and what the risks and benefits might be. Ethical guidelines are developed to specify reasonable and appropriate actions and what protections are needed for all parties involved. Many guidelines also are specified in law to codify protections more formally and to provide recourse in the courts if protections are not adhered to.

16.7.1: Federal Codes and Regulations

Protection of human subjects has a regulatory history in response to the atrocities of the Nazi regime during World War II and the resulting development of the Nuremburg Code (1940). Other codes I have mentioned (e.g., Declaration of Helsinki of the World Medical Association) are in response as well. There are many other codes and they are highlighted briefly.

In 1974, a federal law was passed to create a National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The commission was charged to identify basic ethical principles that underlie research with human subjects and to develop guidelines for research. A report, referred to as the Belmont Report, was produced (DHHS, 1979). The report provided the foundation for many codes and protections that follow. Among the features was establishing guiding broad principles for the protection and care of human subjects, including Respect for Persons, Beneficence, and Justice. The report also elaborated conditions of informed consent and clarification of risks and benefits of the subjects.

In 2001, the U.S. Code of Federal Regulations (CFR, Title 45) elaborated the protections with further details and guidelines (DHHS, 2009a; www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html). Among the contributions was requiring an Institutional Review Board to monitor research at subject protections at research settings (e.g., universities). In turn within the U.S. Department of Health and Human Services, there would be an office to oversee institutions. The oversight commission could call on universities to investigate allegations of violations of rights and provide reports of such investigations. In addition quite specific requirements were set for informed consent and its documentation, institutional processes for reviewing research, and more.

In 1996, the Health Insurance Portability and Accountability Act (HIPAA) was passed as I mentioned earlier in the chapter. Here a key focus was on protecting privacy of health information. Institutions were required to oversee adherence to HIPAA guidelines and to have someone designated as a privacy officer directed to that purpose. The protections include guarding data and identifiable information on-site so that the individual patient is protected. This extent to details such as not leaving patient records on one's desk, not idly chatting about patient information in public places, locking files, and so on.

I have mentioned regulations in brief and defer to the Web sites for further details. The regulations are revised periodically in response to the need, novel applications of research, and violations. From the researcher's perspective, some training usually is required at universities where research is conducted. The training provides familiarity with key tenets and concrete practices (e.g., HIPAA). On a more regular basis, the Institutional Review Board at a university invokes oversight and procedures (e.g., review of research proposals) where the pertinent federal and professional codes are invoked. Meeting the HIPAA requirements at universities occasionally are monitored by visiting research sites to see in fact that patient records are protected.

16.7.2: Professional Codes and Guidelines

The Federal codes have served as the basis of guidelines for research. Many nongovernment professional organizations and agencies involved in delivery of services have their own guidelines. Needless to say, the full range of regulations, codes, and guidelines that govern professional behavior and interactions with the public is too extensive to review here. However, a few are illustrated.

At the most general level, professional organizations provide guidelines to cover a variety of responsibilities. For example, the American Psychological Association (APA, 2010a) has provided an extensive set of principles and guidelines that delineate professional responsibilities in relation to a variety of activities and situations, including research, assessment, interventions (especially psychotherapy), relationships with others (e.g., clients, students), and contacts with the media. Central principles cover obligations in relation to standards of professional competence, integrity, professional and scientific responsibility, respect for the rights and dignity of others, concern for others' welfare, and social responsibilities. Principles from the most recent revision of the codes (APA, 2010a) appear in Table 16.6 to convey the type of statements that are provided. The broad concepts originally in the Belmont Report form a core port of these broad guidelines. These broad guidelines are designed to identify considerations, obligations, priorities, and potential conflicts of interest in relation to the people with whom professionals interact.

At a less abstract level, guidelines cover many specific domains (e.g., use of deception, informed consent). Table 16.7 provides several areas that are included as guidelines for research with human subjects. The domains sample key guidelines to convey the structure and nature of the guidelines. Even moving away from very general principles, the more specific guidelines are still ambiguous and that ambiguity is needed to be applicable to a broad range of circumstances. For example, the principles do not say that deception can or cannot be used. Indeed, the thrust of the principles is to point out the obligations of the investigator and to raise those areas in which caution and deliberation are required. The guidelines point to the considerations included in making decisions about whether a given research project should be undertaken. Although it may be difficult to make decisions in any given case, the overriding concern must be given

Table 16.6: Principles of Psychologists (Abbreviated)

Principles	Description
Principle A: Beneficence and Non maleficence	Psychologists strive to benefit those with whom they work and take care to do no harm. In their professional actions, psychologists seek to safeguard the welfare and rights of those with whom they interact professionally and other affected persons, and the welfare of animal subjects of research. When conflicts occur among psychologists' obligations or concerns, they attempt to resolve these conflicts in a responsible fashion that avoids or minimizes harm. Because psychologists' scientific and professional judgments and actions may affect the lives of others, they are alert to and guard against personal, financial, social, organizational, or political factors that might lead to misuse of their influence.
Principle B: Fidelity and Responsibility	Psychologists establish relationships of trust with those with whom they work. They are aware of their professional and scientific responsibilities to society and to the specific communities in which they work. Psychologists uphold professional standards of conduct, clarify their professional roles and obligations, accept appropriate responsibility for their behavior, and seek to manage conflicts of interest that could lead to exploitation or harm.
Principle C: Integrity	Psychologists seek to promote accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology. In these activities, psychologists do not steal, cheat, or engage in fraud, subterfuge, or intentional misrepresentation of fact. Psychologists strive to keep their promises and to avoid unwise or unclear commitments.
Principle D: Justice	Psychologists recognize that fairness and justice entitle all persons to access to and benefit from the contributions of psychology and to equal quality in the processes, procedures, and services being conducted by psychologists. Psychologists exercise reasonable judgment and take precautions to ensure that their potential biases, the boundaries of their competence, and the limitations of their expertise do not lead to or condone unjust practices.
Principle E: Respect for People's Rights and Dignity	Psychologists respect the dignity and worth of all people, and the rights of individuals to privacy, confidentiality, and self-determination. Psychologists are aware that special safeguards may be necessary to protect the rights and welfare of persons or communities whose vulnerabilities impair autonomous decision making. Psychologists are aware of and respect cultural, individual, and role differences, including those based on age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, and socioeconomic status and consider these factors when working with members of such groups.

Source: Adapted and abbreviated from the American Psychological Association, Code of Ethics (2010a, www.apa.org/ethics/code/index.aspx?item=3).

Table 16.7: Ethical Principles of Psychologists and Code of Conduct

Samples of Codes Governing Research (Section 8 from ethical codes)	Description
8.01: Institutional Approval	When institutional approval is required, psychologists provide accurate information about their research proposals and obtain approval prior to conducting the research. They conduct the research in accordance with the approved research protocol.
8.02: Informed Consent to Research	**a.** When obtaining informed consent as required in Standard 3.10, Informed Consent, psychologists inform participants about: **1.** The purpose of the research, expected duration, and procedures. **2.** Their right to decline to participate and to withdraw from the research once participation has begun. **3.** The foreseeable consequences of declining or withdrawing. **4.** Reasonably foreseeable factors that may be expected to influence their willingness to participate such as potential risks, discomfort or adverse effects. **5.** Any prospective research benefits. **6.** Limits of confidentiality. **7.** Incentives for participation. **8.** Whom to contact for questions about the research and research participants' rights. They provide opportunity for the prospective participants to ask questions and receive answers. **b.** Psychologists conducting intervention research involving the use of experimental treatments clarify to participants at the outset of the research: **1.** The experimental nature of the treatment. **2.** The services that will or will not be available to the control group(s) if appropriate. **3.** The means by which assignment to treatment and control groups will be made. **4.** Available treatment alternatives if an individual does not wish to participate in the research or wishes to withdraw once a study has begun. **5.** Compensation for or monetary costs of participating including, if appropriate, whether reimbursement from the participant or a third-party payer will be sought.
8.03: Informed Consent for Recording Voices and Images in Research	Psychologists obtain informed consent from research participants prior to recording their voices or images for data collection unless: **1.** The research consists solely of naturalistic observations in public places, and it is not anticipated that the recording will be used in a manner that could cause personal identification or harm. **2.** The research design includes deception, and consent for the use of the recording is obtained during debriefing.
8.04: Client/Patient, Student, and Subordinate Research Participants	**a.** When psychologists conduct research with clients/patients, students or subordinates as participants, psychologists take steps to protect the prospective participants from adverse consequences of declining or withdrawing from participation. **b.** When research participation is a course requirement or an opportunity for extra credit, the prospective participant is given the choice of equitable alternative activities.

Table 16.7 (*Continued*)

Samples of Codes Governing Research (Section 8 from ethical codes)	Description
8.05: Dispensing with Informed Consent for Research	Psychologists may dispense with informed consent only: **1.** Where research would not reasonably be assumed to create distress or harm and involves: **a.** The study of normal educational practices, curricula, or classroom management methods conducted in educational settings. **b.** Only anonymous questionnaires, naturalistic observations, or archival research for which disclosure of responses would not place participants at risk of criminal or civil liability or damage their financial standing, employability, or reputation, and confidentiality is protected. **c.** The study of factors related to job or organization effectiveness conducted in organizational settings for which there is no risk to participants' employability, and confidentiality is protected. **2.** Where otherwise permitted by law or federal or institutional regulations.
8.06: Offering Inducements for Research Participation	**a.** Psychologists make reasonable efforts to avoid offering excessive or inappropriate financial or other inducements for research participation when such inducements are likely to coerce participation. **b.** When offering professional services as an inducement for research participation, psychologists clarify the nature of the services, as well as the risks, obligations, and limitations.
8.07: Deception in Research	**a.** Psychologists do not conduct a study involving deception unless they have determined that the use of deceptive techniques is justified by the study's significant prospective scientific, educational, or applied value and that effective nondeceptive alternative procedures are not feasible. **b.** Psychologists do not deceive prospective participants about research that is reasonably expected to cause physical pain or severe emotional distress. **c.** Psychologists explain any deception that is an integral feature of the design and conduct of an experiment to participants as early as is feasible, preferably at the conclusion of their participation, but no later than at the conclusion of the data collection, and permit participants to withdraw their data.
8.08: Debriefing	**a.** Psychologists provide a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research, and they take reasonable steps to correct any misconceptions that participants may have of which the psychologists are aware. **b.** If scientific or humane values justify delaying or withholding this information, psychologists take reasonable measures to reduce the risk of harm. **c.** When psychologists become aware that research procedures have harmed a participant, they take reasonable steps to minimize the harm.

NOTE: This material is a sample of key issues that relate to guidelines for research with human subjects. The ethical codes have many different sections that cover many topics beyond research (e.g., how to resolve ethical issues, advertising, education and training, assessment, and others). Section 8 refers to the codes related to Research and Publication. The full set of codes is readily available (APA, 2010a, www.apa.org/ethics/code/index.aspx).

to the protection of the subject. Indeed, the guidelines specify that as the pros and cons of the research are weighed, priority must be given to the subject's welfare.

16.7.3: More Information on Professional Codes and Guidelines

Guidelines of professional organizations such as the APA overlap but are not to be confused with federal regulations that include law to cover medical as well as psychological research. Beginning in the mid-1960s, the Surgeon General of the Public Health Service required institutions that received federal money for research to establish review committees to consider subjects' rights and to ensure that informed consent was procured for the proposed research. The regulations for research have been revised and elaborated periodically.

> Current federal regulations are designed to evaluate whether any risks to the subjects are outweighed by the potential benefits to them or by the likely benefits to society in light of the information obtained.
>
> In the early 1970s, Congress mandated a special commission to draft ethical guidelines for research with human subjects. The National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research was established to examine research and applications in areas where human rights have been or are likely to be violated. The Commission studied and made recommendations for practices in research with fetuses, prisoners, individuals considered for psychosurgery, and children, all of which raise special issues. These guidelines do not apply to the bulk of research in clinical psychology but are important to mention insofar as they reveal Congress' strong interest in ethical issues raised by research with human subjects. Guidelines continue to emerge to address special topics, such as gene therapy, stem-cell research, but also protection of subjects in general (see NIH, 2011, www.nih.gov/sigs/bioethics/conflict.html).

To examine the risks and benefits and protection of the subject's welfare, research proposals in a university setting are reviewed by a committee that critically examines the procedures and possible risks to the subjects. As noted already, the committee is referred to as an Institutional Review Board in the federal codes to protect subjects (DHHS, 1983). The committee evaluates whether subjects are provided with the opportunity to give informed

consent and made aware of their ability to withdraw consent and terminate participation at any time. Subjects must sign an informed consent form that explains the procedures and purpose in clear and easily understandable language and describes any risks and benefits. Risks are defined broadly to include the possibility of injury—physical, psychological, or social—as a consequence of participation in the experiment.

> Subjects must also be told that they are free to withhold information (e.g., of a personal nature) and to withdraw from the investigation at any time without penalty.

Subjects must be guaranteed that all information they provide will be anonymous and confidential and told how these conditions will be achieved.

Most investigations within psychology include procedures that may be without risk to the subject and hence do not provide problems for review committees to evaluate. Procedures that receive special scrutiny are those projects involving a failure to disclose fully the purpose of the study, deception, the possibility of deleteriously affecting the subject's psychological or physical status, and research involving special populations in which competence to consent is in question (e.g., children). The committee must weigh the merits of the scientific investigation and the advance in knowledge it may provide against possible potential discomfort to the subject.

Ethical responsibility for research cannot be placed solely on the formal review procedures. Ethical guidelines for research encourage investigators to seek advice of others and diverse perspectives to assess whether procedures are warranted that extend beyond minimal risk (e.g., covert observations, discussion of sensitive topics). When weighing scientific merit and the likely benefits of the knowledge against subject rights, the investigator is advised to seek input from colleagues over and above formal review procedures (APA, 2010a). In other words, the ultimate ethical responsibility for the integrity of the research falls to the investigator. There are government resources that provide guidelines, training, and also consultation on critical ethical issues (e.g., NIH Clinical Center, www.bioethics.nih.gov/about/index.shtml).

16.7.4: General Comments

There are many different guidelines that govern research practices and protection of client rights. In this chapter, I have mentioned some of these already (e.g., HIPAA, Declaration of Helsinki, Navajo, and Ho-Chunk Nations). There are scores of other codes of ethical conduct to guide professionals in research. Table 16.8 provides a sample of psychological societies and organizations (e.g., spanning different countries), other professions (e.g., education, statistics), and topic areas that span multiple disciplines including clinical psychology (e.g., addictions, marriage, and family therapy). Some of the connections may seem unobvious. For example, statisticians often are involved in collaborative arrangements with psychologists. There are ethical guidelines for statisticians that have been developed by the American Statistical Association (www.amstat.org/about/ethicalguidelines.cfm). These guidelines underscore important issues regarding the professionalism and professional integrity and the treatment, documentation, analysis, and interpretation of data (e.g., ensure that data conform with any pledges of confidentiality that were made, report on the limits of statistical evaluation of the study along with sources of error, avoid any tendency to slant statistical work toward a predetermined outcome).

Table 16.8: A Sample of Codes of Ethics from Many Professional Organizations in Research Is Conducted and/or Services Are Provided

Sample of Codes of Ethics	Professional Organizations
Samples from Areas Related to Clinical Psychology	• American Association for Marriage and Family Therapy www.aamft.org/imis15/content/legal_ethics/code_of_ethics.aspx • American Counseling Association www.counseling.org/Resources/aca-code-of-ethics.pdf • Association for Addiction Professionals www.naadac.org/code-of-ethics • Human Services Professionals www.nationalhumanservices.org/index.php?option=com_content&view=article&id=43 • National Association for Social Workers www.socialworkers.org/pubs/code/code.asp
Samples of Other Disciplines Often Collaborators with Psychology	• American College of Epidemiology http://acepidemiology.org/sites/default/files/EthicsGuide.pdf • American Education Research Association www.aera.net/AboutAERA/AERARulesPolicies/CodeofEthics/tabid/10200/Default.aspx • American Statistical Association www.amstat.org/about/ethicalguidelines.cfm • National Education Association www.nea.org/home/30442.htm
Samples from Other Countries	• Canadian Psychological Association www.cpa.ca/aboutcpa/committees/ethics/codeofethics/ • Chinese psychological Society www.iupsys.net/images/resources/ethics/china-code-eng.pdf • European Federation of Psychologists' Associations www.efpa.eu/ethics/ethical-codes • Russian Psychological Society http://xn–n1abc.xn–p1ai/en/documents/code_ethics.php • Scandinavia – Nordic Psychological Associations www.iupsys.net/images/resources/ethics/Scandinavian_Code_of_Ethics–English.pdf • Sociedad Mexicana de Psicología www.sociedadmexicanadepsicologia.org/index.php?option=com_content&view=article&id=97

With all the many different guidelines, it is helpful to note that there is considerable overlap. The overarching principles are protection and rights of participants and professional and scientific integrity. In this regard, the guidelines of the APA are comprehensive and often serve as a model for other organizations.

In the various professional ethical codes, the word "guidelines" is used and deserves comment. The word might be distinguished from stronger terms such as rules or mandates. The notion of "guidelines" seems weak—after all, eating the recommended proportions of fruits and vegetables, getting 7 hours of sleep per night, and flossing one's teeth are "guidelines" for good health. I am sure somewhere one of my minor habits (keeping slices of double pepperoni pizza right next to my bed in case I get hungry in the middle of the night) violates a guideline—but who cares? Guidelines are not laws or anything; they are more like advice that somebody went to the trouble to write down. It might be wise to follow them but not mandatory. Perhaps this is so with professional (e.g., APA) "guidelines."

Actually, the term guidelines is partially accurate and partially a misnomer for separate reasons in relation to the ethics of research and practice:

1. Some of the guidelines overlap with state and federal laws and have serious consequences (e.g., prison time, six-figure monetary fines) if they are violated. So, for example, protection of subjects in research and protection of information to ensure confidentiality are central to research guidelines. In addition, federal laws (e.g., HIPAA) in the United States address these issues, and violation of the "guidelines" can result in litigation, punitive action, and more.

   Similarly, some of the guidelines pertain to how professionals represent themselves to the public (e.g., in advertising or listing services). Here too ethical codes and guidelines have been written into various state laws that oversee professionals engaged in practice (e.g., medicine, law, psychology, social work).

   In short, ethical guidelines often overlap with federal and state laws.

   Moreover, if a client in the context of psychotherapy, for example, argues that he or she was treated in a way in which ethical guidelines or commonly accepted professional practices were violated, that too brings to bear legal issues and the courts, even if a very specific law was not available that covered the action. So the guidelines are more compelling than "useful advice you may want to think about."

2. There are several other mechanisms to evaluate adherence to guidelines. Universities and private agencies where research is conducted have federally mandated review boards that consist of a formal evaluation of a given project before it is conducted and training of investigators (doctors, faculty, staff) to ensure that the guidelines are known and followed. The university is accountable as well as investigators for ensuring that the laws and guidelines are followed. Universities can risk loss of research funding from government agencies if they do not properly evaluate research projects and ensure participant protections. At major universities, millions and sometimes hundreds of millions of dollars of research funds could be jeopardized.

A project proposal is submitted and has to convey all the practices that are involved, especially whose to which the participants will be subjected, and what specific practices will be used to protect participants. A project proposal is returned to the investigator for revision or not approved if the criteria are not met and the project cannot be conducted until there is approval. It is in everyone's interest (university, investigator, participant) to ensure that regulations and protections are in place. Investigators rarely can keep up with the changing requirements for research protection, but offices that oversee research at the universities ensure that these requirements are implemented as changes emerge.

Summary and Conclusions: Ethical Issues and Guidelines for Research

Psychological research raises many ethical issues that are intertwined with methodology. Experimental questions and design options, such as those considered throughout the text, do not always make explicit the need to protect the rights and welfare of the subject. Salient issues pertaining to the rights of subjects include deception and debriefing, invasion of privacy, and informed consent. Deception is a major concern when subjects are misguided about the purpose of the experiment and the misinformation may deleteriously affect their beliefs about themselves or others. Deception is infrequently used or indeed permitted in clinical research. If deception is used, subjects must be

informed about the true purposes after they complete the experiment. Providing such information, referred to as debriefing, is designed to erase the effects of deception. However, the effects of deception are not invariably erased. Leaving aside the effects of debriefing, many investigators object to deception because of the relationship it fosters between experimenter and subject.

Invasion of privacy is often protected by ensuring that the responses subjects provide are completely anonymous and confidential. Anonymity refers to ensuring that the identity of subjects and their individual performance are not revealed. Confidentiality requires that the information will not be disclosed to others without the awareness and consent of the subject. In most research situations, anonymity and confidentiality are assured by removing the identity of subjects when the data are evaluated and conveying information publicly (in research reports) only on the basis of group performance. Yet that practice alone may not protect participants sufficiently. Even when the individual cannot be identified, it is possible that large segments of society (e.g., particular ethnic and cultural groups, a restricted geographical setting) will be adversely affected by unwittingly impugning characteristics of the group or residents of the community.

Informed consent is a central issue that encompasses many ethical concerns and means of protecting subjects in experimentation. Informed consent requires that the subject willingly agree to serve in an experiment and be fully aware of the procedures, risks, and benefits when making the choice to participate. Procuring and interpreting informed consent is not entirely straightforward because the subject must be competent to provide consent, know the relevant information to make a meaningful choice, and consent completely voluntarily. Whether these criteria are met in any given case often is a matter of debate. For minors, assent is obtained, which consists of agreement of the child to participate in the project. Assent does not replace obtaining informed consent from a legal guardian.

The many ethical issues raised in research have prompted guidelines and regulations designed to protect the rights of individual subjects. The guidelines apprise investigators of their obligations and the priority of ensuring protection of the subject at all times. The guidelines do not necessarily rule out practices that might be objectionable (e.g., deception). However, the onus is upon the investigator to show that there are likely benefits of the research and that these require a departure from full disclosure of the purposes and procedures. Regulations are more demanding, and they put into policy and law precise protections that subjects are afforded. This chapter mentioned some of these federally mandated requirements (e.g., HIPAA) as well as additional protections (e.g., Certificate of Confidentiality).

The concern over protection of participants is an international focus, and individual countries have guidelines. Also many guidelines span multiple countries (e.g., Declaration of Helsinki).

Within the United States too, there are multiple guidelines. I mention that Native American nations, some of which have regulations that guide research (e.g., Ho Chunk, Navajo). For the individual investigator, the research setting often has multiple resources to convey explicitly what the requirements are that need to be met to conduct research.

This chapter has focused on ethical issues raised in the context of research and responsibilities in relation to protection of the rights of participants. The broad area of I have addressed still is only half of the story, i.e., set of ethical responsibilities. The other half pertains to scientific integrity and many facets of behavior of investigators. These too have guidelines, regulations, and professional and legal consequences, although they often are part of a broad set that applies to both ethical issues and integrity. Scientific integrity is treated as a separate chapter because the issues warrant their own elaboration, cautions, and guidance.

Critical Thinking Questions

1. What is deception in research? Give an example (hypothetical or real) when it probably is not much of a concern and another example when it would be ethically inappropriate and maybe even illegal.
2. What is invasion of privacy as an ethical violation? Give an example of something that would qualify.
3. What are informed consent and assent? How are they different?

Chapter 16 Quiz: Ethical Issues and Guidelines for Research

Chapter 17
Scientific Integrity

Learning Objectives

17.1 Recognize some of the inherent value systems that guide scientific integrity

17.2 Report the general guiding principles that encourage scientific integrity

17.3 Express some of the concerns in maintaining scientific integrity

17.4 Recall scientific ethical codes as applicable to authorship and allocation of credit

17.5 Review the advantages and disadvantages of sharing of materials and data of scientific work

17.6 Examine how conflict of interest may emerge in scientific research

17.7 Identify instances that cause breaches in scientific integrity

17.8 Determine remedies and protections to safeguard ethical interests of the subjects of statistical research

We have distinguished ethical issues as the responsibilities of researches in relation to participants from scientific integrity as those responsibilities related to the standards and obligations in conducting and reporting research. These can be distinguished at the margins (e.g., ethical issues related to ensuring that informed consent is voluntary for each subject vs. integrity issues such as plagiarizing or fabricating data). Yet ethical issues and integrity overlap in core values (e.g., transparency, honesty) and in how one represents one's work (e.g., as we describe the project to the participants and as we describe our work to other professionals and the scientific community).

Even with the overlap, it is quite valuable to treat scientific integrity on its own because of the host of issues that are raised. Also, integrity is not treated the same ways as are ethical issues we have covered. In the protection of individual subjects, various review panels at universities (e.g., Institutional Review Board, HIPAA privacy officers) evaluate research in advance of running a study. Before a study can be initiated, proposals and procedures are vetted for approval to ensure that planned protections are in place. Then annually, or sooner if something unexpected occurs, an investigator usually has to report to the review board how the project is going and whether any untoward events (e.g., side effects, subjects withdrawing) have occurred.

In contrast, most of the scientific integrity issues are out of view of oversight committees and often emerge during or after the study. For example, if the data are faked or massaged or if several measures included in the study never make it to the final analysis, this is not easily picked up by any committee.

> Within a university or institution, integrity issues often arise long after a study has been completed.

Only when accusations are made by someone about a lapse in treatment integrity and the university administration has reason to suspect foul play does the matter arise. At that point, there is likely to be a committee convened in the university to investigate the accusations.

In decades past, questionable research practices by an investigator might have made the local news, may have led to a university investigation, and some settlement months later would have been reached. Now that same event is likely to be picked up nationally or internationally, in light of the Internet, social media, and news agencies, all of which can communicate information more extensively and rapidly than ever before. Also, now people can "weigh in" and provide comments at the end of a story, and soon the commentators are talking to each other, add a few tweets, stir in a few well-placed blogs, and an international news event is born. Yet apart from questionable research practices, science in general

is in the news more frequently (e.g., with breakthroughs in discovery and technology), and these circulate in the same way through news and social media. The communication methods and increased involvement of the public have increased visibility of lapses of scientific integrity and greater and the increased accountability to which this leads. As part of that universities are scrutinized to be accountable for what they are doing to investigate and perhaps what they did not do to begin with that may have fostered the problem. The attention to science including lapses of integrity and stunning discoveries overall is a benefit in large part because that attention is consistent with core values of sciences and specific practices designed to reflect and maintain these values.

17.1: Core Values Underpinning Scientific Integrity

17.1 Recognize some of the inherent value systems that guide scientific integrity

Scientific integrity begins with core values of science that underlie the approach and how one "does" science. Those values are not about methodology per se but have critical implications for methodological practices. These core values are nicely previewed by an eminent scientist, Albert Einstein, who noted, "Many people say that it is the intellect which makes a great scientist.

They are wrong: it is character" (National Research Council, 2002, p. 16). And by character we are referring to the values and standards one has, invokes, and follows in conducting science. We are talking here about character of scientists and the characteristics of science. As examples, transparency and honesty are two such qualities just to preview our discussion. And these terms seem too general to be of much value in guiding ethics or scientific integrity. Yet, they are essential because one cannot anticipate all of the practical situations or issues in which issues of scientific integrity will emerge. Hence it is important to be able to refer to basic values to guide new situations. Second are the many concrete practices that can be used to foster, promote, maintain, and enforce scientific integrity. These span education and training, publication of articles, and efforts to catch lapses of integrity. Let us begin with the abstractions as a context for discussing key issues and concrete problems of scientific integrity.

Scientific integrity begins with several core values. They include the "character" of individuals in the way Einstein mentioned but also as defining features of science. Table 17.1 provides these core values. They are abstract, but each becomes much more concrete as we move forward in the chapter.

The concepts may seem obvious and self-evident at first blush and perhaps hardly worth making explicit. Yet, as we see later in the chapter, there are many ways in which the values and practices derived from them come into conflict and have nuances.

> There are conflicts and nuances of some core values in everyday life too outside the context of research, although not often discussed.

Table 17.1: Core Values and Underpinnings of Science

Value	Definition
Transparency	Openness of what one is doing and has done. Procedures, methods, and findings are not to be restricted, secret, or obscured. Rather, the effectiveness of science in developing and revealing knowledge depends on one's peers and the public to be able to see what was done and how it was done. This is required for replication of findings, but transparency as a value goes well beyond replication. Findings and procedures must be available to others in principle and practice.
Honesty	The accumulation of knowledge depends on accurate rendition of all actions that are part of research. Acts of commission or omission that misrepresent any facet of conducting and reporting on a study undermine science.
Accountability	The scientist is responsible for actions and activities related to planning, executing, supervising, and other activities involved in the research process. This means that the scientist has special responsibilities in adhering to the standards, ethics, and other core values of science and can be held accountable for violations. Accountability is not only to the standards of science but also to multiple constituencies (colleagues and collaborators, institutions, students, and the public at large). To all such parties, the core value is to represent oneself, one's discipline, and one's research with the other values noted here (transparency, honesty).
Commitment to Empirical Findings	Scientists focus on and yield to empiric evidence as the arbiter of knowledge. This is distinguished from other ways of knowing, such as arguing from authority or from faith. There is no implied criticism of other ways of knowing, but rather clarification of the commitment of science to the accumulation of empirical knowledge. Invariably there are disagreements, debates, and contrasting theories about a particular set of findings, but these are not the issue. In the end, findings, replicated findings, and the accumulation of knowledge are what science uses as the basis of knowledge.
Conflict of Interest	As individual scientists may have allegiances, commitments, or that could influence or give the appearance of influencing the findings. For example, the scientist may have invented a procedure or product used in research or clinical care and has a commercial interest in its success. Also, the investigator may have received grants or consulting fees from a company that has a definite position or goal (e.g., supporting the effectiveness of a procedure or medication). Conflict of interest may influence findings that are obtained and reported. A scientific value is trying to avoid conflict of interest and reporting such influences or their appearance.
Commitment to the Public Interest	Science serves the public. Our knowledge is rooted in an effort to understand our world (and other worlds) in a way that can ultimately improve life in all of its forms. Apart from goals, ultimately much of research is funded or otherwise supported (condoned, respected) by the public.

For example, most parents probably favor honesty and model behaviors that reflect honesty. *Would parents with that value systematically lie to their children?*

What a horrible thought! Well, not really horrible. They knowingly lie to their children regularly and predictably.

Many people (probably most readers) early in life were led to believe in Santa Claus or the Tooth Fairy as veridical individuals with a specialized agenda and who visited one's house. (My deep apologies if this text is the first news any reader may have that Santa and Toothy are not real or the way you believed—I am just the messenger.) Add to those parent fostered falsehoods that comments from our well-trusted pediatricians who once in a while said, "This won't hurt"—surely they too must have known. As I mention these, it is easy to say, "Well, those are not really lies or if they are they do not 'count.'" But this is the issue that we will come to in more serious contexts. In a more serious context, there are conflicts of values and suspension of some of these (including honesty) in special circumstances. Thus, there are nuances that make conducting science and adherence to values challenging. The challenges continue to emerge because of novel research methods and ways of utilizing data, and these often chart new ground with challenges for both ethical issues and scientific integrity.

17.2: Ethical Codes Related to Scientific Integrity

17.2 Report the general guiding principles that encourage scientific integrity

Ethical codes of the APA were presented in relation to the treatment and protection of participants in research and the obligations of the investigator (e.g., deception, informed consent). The codes begin with broad principles that address beneficence and nonmaleficence, fidelity and responsibility, integrity, justice, respect for people's rights and dignity (please see Table 16.6). Those general principles encompass scientific integrity as well. Several specific codes move from the general case to issues that directly address scientific integrity.

Table 17.2 provides specific topics that relate to scientific integrity. The guidelines in the table are important to

Table 17.2: Ethical Principles of Psychologists and Code of Conduct

Samples of Codes Governing Research (Section 8 from ethical codes)	Description
8.10 Reporting Research Results	**a.** Psychologists do not fabricate data. **b.** If psychologists discover significant errors in their published data, they take reasonable steps to correct such errors in a correction, retraction, erratum, or other appropriate publication means.
8.11 Plagiarism	Psychologists do not present portions of another's work or data as their own, even if the other work or data source is cited occasionally.
8.12 Publication Credit	**a.** Psychologists take responsibility and credit, including authorship credit, only for work they have actually performed or to which they have substantially contributed. **b.** Principal authorship and other publication credits accurately reflect the relative scientific or professional contributions of the individuals involved, regardless of their relative status. Mere possession of an institutional position, such as department chair, does not justify authorship credit. Minor contributions to the research or to the writing for publications are acknowledged appropriately, such as in footnotes or in an introductory statement. **c.** Except under exceptional circumstances, a student is listed as principal author on any multiple-authored article that is substantially based on the student's doctoral dissertation. Faculty advisors discuss publication credit with students as early as feasible and throughout the research and publication process as appropriate.
8.13 Duplicate Publication of Data	Psychologists do not publish, as original data, data that have been previously published. This does not preclude republishing data when they are accompanied by proper acknowledgment.
8.14 Sharing Research Data for Verification	**a.** After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release. This does not preclude psychologists from requiring that such individuals or groups be responsible for costs associated with the provision of such information. **b.** Psychologists who request data from other psychologists to verify the substantive claims through reanalysis may use shared data only for the declared purpose. Requesting psychologists obtain prior written agreement for all other uses of the data.
8.15 Reviewers	Psychologists who review material submitted for presentation, publication, grant, or research proposal review respect the confidentiality of and the proprietary rights in such information of those who submitted it.
3.06 Conflict of Interest	Psychologists refrain from taking on a professional role when personal, scientific, professional, legal, financial, or other interests or relationships could reasonably be expected to: **a.** Impair their objectivity, competence, or effectiveness in performing their functions as psychologists **b.** Expose the person or organization with whom the professional relationship exists to harm or exploitation

NOTE: This material is a sample of key issues that relate to guidelines for research with human subjects. The ethical codes have many different sections that cover many topics beyond research (e.g., how to resolve ethical issues, advertising, education and training, assessment, and others). Section 8 refers to the codes related to Research and Publication. Those related to ethical issues and the investigator's relation and obligations to the participants are part of Section 8. In that same section are issues related to scientific integrity and reproduced here. Added in the present table is the code from another section (Section 3 on Human Relations). This section includes Conflict of Interest, which extends beyond research and publication. The full set of codes is readily available (APA, 2010, www.apa.org/ethics/code/index.aspx).

note to convey the obligations of scientists. They are not designed to get into the practical details about how they are to be implemented. Nevertheless they provide a valuable point of departure for us as we do take up many of the specific issues.

17.3: Critical Issues and Lapses of Scientific Integrity

17.3 Express some of the concerns in maintaining scientific integrity

Scientists are not immune to error, deception, and fraud in their work. Historical accounts of science provide a long line of examples in which scientists have made major errors in recording or interpreting their work, have tried to deceive others about their findings, and have altered or faked their data (see Miller & Hersen, 1992; Moore, Derry, & McQuay, 2010).

17.3.1: Fraud in Science

The initial scientific integrity issue mentioned in Table 17.2 focuses on the accuracy of scientific reporting. The language is gentler than my heading of fraud. The matter of accuracy of reporting can mean a few different things. Let us begin with error, move to fraud, and then to other ways in which findings are distorted. The distinction between error and fraud in research is major.

> *Errors refer to honest mistakes that may occur in some facet of the study or its presentation.*

The processes entailed by collecting data, scoring measures, transcribing and entering data, and publishing raise multiple opportunities for error. To err is human; to err frequently is careless, and to err in a particular direction (e.g., to support a hypothesis), at the very least, is suspicious. The investigator has the responsibility to minimize error by devising procedures to check, monitor, and detect errors and then to rectify them to the extent possible. When errors are detected (e.g., as in published reports), investigators are encouraged to acknowledge them as quickly as possible. Often journal publications include isolated notes (referred to as errata) where corrections can be written in the same outlet in which the original paper appeared or a statement is made that the original finding is retracted (see Fang, Steen, & Casadevall, 2012).

Errors are important and it is critical not to be dismissive about them. For example, one respected review (Cochrane Collaboration) examined the impact of methods of communicating with patients in ways that would promote changes in health-related behavior.[1] The review concluded that the interventions were harmful (see Moore et al., 2010). Although the result made it to the media, almost immediately, the authors realized they entered their data in a way that was incorrect. The exact opposite conclusion was the one supported by the data, and the finding was corrected with a retraction. Unfortunately, retractions (journal disclaimers about a previous finding) did not make it into the news, and in fact this retraction was not cited very much either. Errors can make a difference and from the standpoint of public impact on the findings themselves, fraud and errors share at least one feature, namely, that the conclusions are wrong. To the extent that the finding relates to some facet of everyday experience (e.g., mental or physical health, child rearing, elderly care) makes the incorrect information potentially harmful in concrete ways.

Errors usually are not considered as part of lapses in scientific integrity. Fraud certainly is because it is not carelessness or human error.

> *Fraud in science refers to explicit efforts to deceive and misrepresent.*
>
> Of all scientific integrity issues, fraud is the most flagrant because it undermines the foundations of the entire enterprise of scientific research.

Although fraud is not new in science, recent attention has focused on deliberate efforts of researchers to mislead colleagues and the public. Dramatic instances have come to light in which critical procedures, treatments, or potential breakthroughs could not be replicated or were known by one's colleagues to reflect explicit attempts to misrepresent the actual findings. It is important to note that issues of fraud encompass all of the sciences. The examples selected from psychology are critical to the topic of this text, but it would be misleading to imply that the issues and violations apply uniquely to psychology (e.g., see Fang et al., 2012).

Some examples were mentioned in passing in the context of negative findings and replication. I present these in more detail here because they are central to scientific integrity, have had enduring impact on the public and scientific communities, and in one case turn out to be a matter of life and death. One example is the study suggesting that a commonly used vaccination for children (one vaccination measles–mumps–rubella) may cause autism (Wakefield et al., 1998). The study is now "old," but the issues and consequences are evident today. Actually, this was not a controlled study but a report (case studies) of 12 children who allegedly had been functioning normally but after vaccination lost the acquired normal skills and showed behavioral symptoms of pervasive developmental disorder. The disorder was attributed to the vaccination. This led to:

- Public airing of the findings
- Deep concerns among parents, books, television talk shows, news specials, claims of government conspiracy

to hide the fact about what was putatively known about vaccinations
- Celebrities taking up the cause that vaccines were the culprit
- Death threats on and against leaders of drug manufacturers (of the vaccine)
- Congressional Hearings, an Institute of Medicine Report, endless efforts to study and replicate the finding (e.g., Deer, 2011; Langan, 2011; Sugarman, 2007)

An initial Institute of Medicine panel of experts evaluated the findings and subsequent research to conclude there was no evidence supporting the connection of vaccination and autism (Stratton, Gable, Shetty, & McCormick, 2001). Later and updated reviews now encompassing hundreds of studies supported the same conclusion (DeStefano & Thompson, 2004; Miller & Reynolds, 2009). While wrenching parent trauma and replications were going on, finally years later the original article was identified as a fraud. The authors purposely misrepresented the data. Autism could not be traced to the vaccines at all, and in fact children in the original report had problems before being vaccinated. There was no clear connection between vaccination and autism—the data were faked (Editors of *the Lancet*, 2010).

It might be fair to say that all fraud hurts science and the public in multiple ways such as undermining public trust and support for research, generating cynicism that "knowledge and truth" really mask individual motives and ambitions, and wasting intellectual and financial resources because a finding may lead other scientists to pursue lines that have no basis. Yet not all fraud literally harms the public. This one did because many children have died and continue to die because of the fraudulent finding and attention it received. The findings promoted large-scale suspicion about the dangers of vaccinations despite endless retractions, expert panels, and evidence that the original finding was bogus. Vaccination rates in the United Kingdom, United States, and other countries declined and the diseases they were designed to prevent have increased (Gross, 2009). This is what has now made the scientific fraud an issue of life and death. Even currently as vaccination rates have increased, many parents continue to avoid vaccinations. How to rebuild public trust?

As for Wakefield, a formal review of his work focused on his conduct of the research rather than the findings. After 2½ years of investigation, he was accused of conducting invasive medical tests (blood samples), did not seek appropriate consents, had a conflict of interest (being compensated to advise lawyers about the harm caused by vaccines), not qualified to do this research, and more—guilty of a total of 30 charges and that led to removal from the ability to practice medicine (Triggle, 2010). I mentioned that journal retractions and great publicity did not reverse the impact on vaccinations. Organizations, conferences, anti-vaccination movements, and Web sites help maintain the fear of vaccinations. Wakefield moved (to the United States) and promoted a book that helps maintain the fear of vaccinations. And there are many fans who follow and endorse his work, and many parents with wrenching decisions to make—should I have my child vaccinated to prevent some diseases but risk autism (Dominus, 2011). It does not matter in some way that the risk of autism has been dispelled.

17.3.2: More Information Regarding Fraud in Science

As an additional example, a relatively recent case of fraud by a psychologist received widespread attention. A prominent social psychologist from the Netherlands (Diederik Staple) was accused and found guilty of faking over 50 publications. Over a period of 15 years, several publications were published often in prestigious journals; the studies and their findings were later found to be fabricated. Dr. Staple had worked at three different universities in the Netherlands, and each mounted a committee to evaluate his work. The work culminated in a one final report (referred to as "Flawed Science" in a publically available document) (see Levelt, Noort, & Drenth Committees, 2012). Unequivocally, fraud was uncovered in scores of his publications where he simply faked the data without running subjects or added and changed numbers as if they were from data collected for the studies. In light of the evidence, Staple was suspended from his university.[2]

These examples both involve individuals and hence might imply that fraud occurs in the context of individuals and in the privacy of their own labs, with the lights turned down, and the blinds down. That may be true, but it is important to note in passing not always. Fraud can occur at a larger scale. For example, in the recent news are allegations about a study in Japan looking for early signs of Alzheimer's disease and to use that information to develop medicines and treatment methods (Watanabe & Aoki, 2014). The research involves neuroimaging and blood tests to determine the relationship with various symptoms of the disease.

> The study involved 38 medical institutions and received funding from health and education ministries of the government and 11 drug companies.

The allegations include rewriting of the data or requests from the data center overseeing the research to other sites to redo the data on multiple occasions, using subjects (too ill) who were inappropriate for the study, and more. As I write this, the outcome does not look

great but this is the allegation stage and investigations are beginning.

Fraud in science on any scale is egregious in its own right and apart from any sweeping consequences that might be identified. And yet, as one of the examples (Dr. Staple) illustrates, public distrust can be enormous and go well beyond the confines of the original fraud (e.g., vaccinations). Extensive attention to the case in print and online media and professional journals brought worldwide attention to the work on vaccinations and autism. The reactions included accusations that social psychological research, psychological research, and perhaps all social science research should be mistrusted. Yet, the entire matter is much broader and shakes the very foundation of all of science. As with the case on the faking of an association of vaccines and autism, the public trust was easily lost and widely generalized. It is not so easy to rebuild that trust. From the standpoint of the media, all the findings that might be cited to build trust or to focus on legitimate accomplishments (e.g., effective treatments of psychological or medical conditions) understandably and lamentably are not as newsworthy as the scandals.

A few points are worth underscoring:

- Fraud is inherently against the core tenets of science. That point means that independently of any fallout, it is something that violates all we are doing.
- Second and distinguishable are the deleterious consequences. Undermining public trust means that science in general will be under suspicion.

Occasionally, in relation to health (e.g., vaccination, participation in health care system) raising suspicions can harm people. Child vaccinations worldwide, for example, save millions of lives each year. Many who are suspicious of vaccinations and probably suspicious of the "medical establishment," as if it were one unified position more generally may be less likely now to utilize treatments that will help them or their children. The distrust for many has gone well beyond vaccinations.

How prevalent is fraudulent behavior?

This is inherently difficult to answer—fraud is hidden and only the "failures" at hiding are detected. Also, fraud can take many forms and a single percentage would not capture or represent that very well. Yet, it is worthwhile to provide some information. In a review involving several other studies, one of which included approximately 12,000 scientists, survey data revealed that 2% of those who responded reported fabricating, falsifying, or altering their own data (Moore et al., 2010). In addition 15% said they know of other people who did so. Estimates of fraud in the review article were placed at between 1% and 13% of the researchers. Again, it is difficult to consider any number precise but still instructive that survey data reveal even numbers as 2%. There are thousands of studies published each month, if one considers all the sciences, and no doubt hundreds of studies each month at the minimum if one restricted this to all of psychology. Actually, we can be a little more precise; in a given year drawing from the Web of Science, over 1.4 million papers are published yearly encompassing over 250 disciplines (Uzzi, Mukherjee, Stringer, & Jones, 2013). Are approximately 1% and 13% of these fraudulent? That would be a disaster and would underscore even more the essential role of replication.

No single or simple cause of fraud is likely to be identifiable. No doubt several influences can conspire to produce fraud, including:

- Pressures of investigators to publish for recognition
- Career advancement, to produce positive results (an effect), to make critical breakthroughs
- To obtain funding for seemingly promising avenues of research

For many academic researchers, often in medical schools, grants are the bases for their salaries and not obtaining a grant can lead to a salary cut. Protections to counter fraud include:

- Training in the central values of science
- Emphasis on the importance of the integrity of the investigator in advancing science generally, repeated opportunities in which an individual's work is subjected to review
- Evaluation by one's colleagues (e.g., as part of review processes in advance of the study, when the report is submitted for publication)
- Public access to data records
- Efforts to replicate research

Apart from sanctions within the professions, there are legal consequences of fraud as well. For example, the use of public funds in research and conduct of research within institutions brings to bear specific laws regarding fraud and oversight boards to monitor scientific integrity, to investigate allegations of fraud, and to pursue through the courts culpability and prosecution. The sanctions and consequences can be personally and professionally devastating. Indeed, being accused and later completely acquitted in one well-publicized case of a researcher whom I knew well had very negative consequences from which it was difficult to recover. (I do not mention the name for risk of continuing the unwarranted stigma with which the name was associated during the accusation and trial period.) Notwithstanding the multiple factors to protect against fraud and to invoke various sanctions, the key protection remains. There must be an ethical commitment and responsible behavior

on the part of the individual investigator in conducting studies, reporting data, and preparing reports.

> In all these activities, an investigator ought to be as honest and objective fashion as possible and to train those working under one's charge (e.g., students) in these standards and practices as well.

I will say more about remedies later after considering different types of lapses in scientific integrity.

17.3.3: Questionable Practices and Distortion of Findings

Fraud involves flagrant attempts to mislead and usually is reserved for plainly fabricating, inventing, and distorting data. I believe it is useful to include a category of questionable practices in which the reporting of the procedures and findings of a study is incomplete, incorrect, and misrepresented. I am referring that to questionable practices but still involve distortions of the data that too might be judged as flagrant.

In any investigation and its write-up for publication, multiple decisions are made as to how the data will be analyzed, summarized, and reported. There are multiple opportunities to select among many options so as to yield significant findings and present a very slanted or incomplete picture of the study (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). In Table 17.3 I have listed several opportunities and decision points. In each of these, the investigator may have many options

Table 17.3: Questionable Practices and Decision Points That Can Distort the Findings

Decision Point and Practice	Elaboration
Data Analyses	Many different statistical tests can be carried out to evaluate the hypotheses, and the tests do not necessarily lead to the same finding or conclusion. After multiple analyses, the one(s) that "worked," i.e., showed statistical significance, may be selected and reported without sharing information about all of the analyses used to reach that point.
Including or Excluding Outliers	Outliers are individuals whose data are extreme on a measure as defined by departing from the mean value. An outlier might be defined as one whose score is 2, 3, or more standard deviations above the mean. The investigator may or may not exclude subjects who are outliers and may vary in the cutoff (how far from the mean) to define outliers. The decisions (exclude or not; at what cutoff if excluded), if not made in advance of the study, might be based on the data analyses. That is, the investigator may explore the impact of the decision rule on the statistical significance of the findings. The findings may or may not be statistically significant or vary in strength as a function of whether subjects are excluded.
Including or Excluding Subjects Who Varied in Their Responses to the Manipulation Check	In some studies, a measure (e.g., questionnaire) is used to see if the subjects experienced, recalled, or perceived the experimental manipulation. For example, the goal may have been to produce positive or negative moods in different groups and a brief set of questions may be provided to see if the appropriate moods were achieved as measured by those questions. A decision has to be made about what to do if individuals did not show that they perceived the manipulation. This is a problematic issue for a variety of reasons (e.g., manipulation check measure is not usually a validated instrument; a manipulation can still be very effective no matter what the manipulation check shows). The investigator may exclude some subjects based on their performance on the manipulation check. The bias comes from analyzing the data by variations of who is included or excluded based on that manipulation check and some cutoff used to define who perceived that manipulation and who did not.
Selecting Control Variables (covariates)	In any study, there will be individual difference variables such as age, ethnicity, sex, socioeconomic status as the more common ones but others that may seem pertinent to the study (e.g., level of stress, history of drug abuse). These variables can be "controlled" or rather taken into account in data analyses in many ways by counting them as covariates (as one among many statistical options). Whether and how these are considered in the data analysis can lead to differences in findings. Different data analyses may be selected after the investigator explores many options (controlling for covariates or just one or none).
Sample Size and Peeks at the Data	We think of the sample size selected in advance of a study, ideally based on evaluation of statistical power. Yet, as the data are collected, investigators occasionally analyze the results before the full set of subjects has been run. The checks may see whether the results are "coming out" the way they were predicted. The biases come from checking repeatedly—which increases the likelihood of a chance finding from multiple statistical tests and from deciding to stop running subjects or to continue based on whether the findings were significant.
Selecting among Dependent Variables	A study may include many different measures to evaluate the impact of the manipulation. The investigator may only report those that showed the effect. We have no idea of the extent to which the data reported represent the full set of measures and many analyses.
Selective Use of Subscales	Some measures yield total scores but also may have many different subscales (e.g., to measure kinds of social support, or stress, or facets of a relationship). Bias enters in if select subscales are included or omitted based on what they show. Two of 10 subscales may show an effect and only those are included in the method and results section.
Selective Use of Experimental Conditions	A study may have multiple groups that receive different conditions. In the final reporting of the study, some groups may be excluded or may have been combined. The decision may be made based on the yield from the data analyses, i.e., whether the results are statistically significant based on what groups are included.
Not Reporting or Writing Up Results	An investigator may have conducted many studies on the topic, and some or most of these may not have shown the effect. Publishing the one study that does show the effect distorts the set of findings across studies. The difficulty in publishing negative results is a disincentive for writing up such studies.

about how to proceed. There are no firm guidelines for many of the specific decisions listed in the table. The available options for a given decision could be defended. What raises the problem is that the investigator may make the decision based on looking at the various options, seeing which ones lead to the expected conclusions, and then reporting just those options.

Several of the points in this table have been discussed by others (e.g., John et al., 2012; Simmons et al., 2011).

17.3.4: More Information on Questionable Practices

Consider a few examples that expand on the table. Let us begin with the data analyses because these can encompass many questionable practices under one rubric. In our study, we have a fairly good idea of our hypotheses. Also, as the data come in they are automatically stored on a database. Software programs make it easy to analyze the data, so perhaps we should take a peek to see how the study is going even though we have many more subjects to run. Let us say, one third of the subjects is run and we peek to see how the data are coming in. We find no statistically significant effects, so we decide to continue to run subjects. The next check we do might be after two thirds of the number of subjects we planned to run. We complete various statistical analyses again, but this time we find the predicted effects. That is, now the results are statistically significant. We decide to stop on the spot. Why collect further data when we have shown the predicted effect was found?

What do you think?

The answer is that multiple checking on the data in this way increases a likelihood of chance finding and biased decision making (stop if an effect, continue if no effect) capitalizes on that bias (Francis, 2012).

17.3.5: Another Data Analysis Point

Let us say we gained control of our statistical voyeurism (peeking where we shouldn't). In our next study, we run all the subjects as originally planned and do not peek at how the results come out until that process is completed. Now we analyze the data. With statistical package software programs as user-friendly as they are, it is easy to analyze the data in many different ways with just a matter of clicks. Usually there are many suitable ways to analyze the data to test a hypothesis. Not all of the data analyses yield similar results and a given "finding" may be strong, weak, or disappear depending on the analysis that is selected. A decision may be reached as to what data analyses to use or report based on what the separate ways of looking at the data actually show. Obviously this is a bias.

Clearly, the practices I have outlined introduce serious bias and are questionable. Other points in Table 17.3 relate to data analyses too, and so let us combine some of those with the above. We may not just do the different analyses as noted above, but repeat each one of them based on excluding some subjects that we define as outliers and excluding others because of how they responded to some manipulation check. Also, we could be doing this while making early peeks at the data before all the subjects are run as I already described. Now we have multiple analyses, being explored with different sets of subjects (depending on who if anyone is excluded), and are doing all of this at different points (early peeks at the data). We may report findings as significant at $p < .05$, but in fact with so many tests and our plucking those we found as significant, that is not the real p level. The chances become remarkably high that we find something and that the finding is one easily explained by chance from multiple tests. Rarely do investigators report the process leading to how a particular test was plucked out and emphasized.

Incomplete reporting may extend beyond the data analyses and pertain to the measures. There may be many measures that were used to evaluate the hypotheses. The results may be different based on which measures are presented in the final report. It is very likely that some measures may not be reported at all or perhaps only portions of measures (one or a few subscales). Or perhaps some of the trials (early, mid, or late in the session on a cognitive task or neuroimaging results) were more revealing or interpretable than others. These more revealing trials may be the ones reported or emphasized. In the final results, the reader may not be informed about the decisions about what measures were presented and all the options that were considered. In the discussion of replication, I emphasized how this selective reporting can lead to chance findings and findings not likely to be replicated. In this chapter, the same behavior is viewed in a light of a different emphasis, namely, scientific integrity. Not reporting the measures and data analysis is deceptive. Complete transparency would be reporting all measures, all analyses, and all efforts to look at the data to see what emerged.

The example illustrates the point. Multiple decisions made in a study represent points where questionable practices can readily emerge and distort the process of what was done and the full outcomes of the analyses. Table 17.3 provides a summary of key points and how they can mislead. The experimental design and the written article reporting on that design and findings may be linear, straightforward, and simple. Yet, behind the scenes more was done. Table 17.4 gives a rendition of what I mean by a linear report or rendition of the study (upper portion) and

what it may have taken for the investigator to get that rendition (lower portion). The nonlinear rendition involves what often amounts to "trial and error" evaluation of the results leading to selection of the one that led to a significant finding.

Table 17.4: Linear and Straightforward Report of a Study and Circuitous Path to Get There

Linear Reporting of What Happened in the Study	Circuitous Path to Get to the Linear Reporting
• I predicted that this would happen • I used these measures • I selected this many subjects • I comprised these groups • I did these data analyses • I have these findings to report	In fact, the study may be more like this: • I predicted that this would happen (and a lot of other predictions that I dumped because they did not come out) • I used these measures (and many other measures but dumped them for various reasons) • I selected this many subjects (actually a lot more, but tossed a few outliers and others for various reasons) • I comprised these groups (but had one more group that I had to toss because their results obscured things) • I did these data analyses (among a seemingly endless set of options, with and without controls for possible confounding variables, with and without some of the subjects I dumped, and more—don't ask) • I have these findings to report, which is the tip of the iceberg

Overall, the discussion and Table 17.3 describe questionable practices. Solutions have been identified. They include requiring complete reporting of all that was done (all measures, all analyses) and encouraging journal editors, reviewers, and authors to publish negative (no difference) results (e.g., Nosek, Spies, & Motyl, 2012; Simmons et al., 2011). Many of the practices can be handled by specifying in advance of a study how decision points will be handled. That is:

- Making decisions of how many subjects will be run (no multiple peeks at the data)
- Who will be excluded if anyone (e.g., outliers, manipulation check failures)
- What will be the likely variables to be controlled in the statistical analyses (e.g., covariates)

I mentioned previously that in advance of a study funding agencies may ask the investigator to specify several critical decision points before the study is run (see ClinicalTrials.gov). Also, when manuscripts are submitted for publication, journal guidelines may require the investigators to provide information on exactly what measures were used, how they were evaluated, and so on (DeAngelis et al., 2005; Laine, Goodman, Griswold, & Sox, 2007).

Some of the questionable practices and biases stem from exploring the data, i.e., conducting analyses that were not originally planned or specified. It is quite fine to explore one's data. Indeed, one should look carefully for nuggets that might be mined for the next study or line of research. This is different from searching for significance and engaging in questionable practices as outlined in the table as a way of testing hypotheses.

Some of the pressures for the questionable practices stem from the publication bias that favors publication of results that show statistical significance. As we have discussed, two broad influences with renewed attention and interest now are in replication research and publication of negative findings. Both of these alter the incentive structure a bit for finding that correct arrangement of subjects, decisions, statistical tests, and measures so that the hypotheses "were supported."

17.3.6: Plagiarism

Plagiarism refers to the direct use and copying of material of someone else without providing credit or acknowledgment.

This can include words or ideas that one attributes from another person that one attributes to oneself. One can see in the ethical codes of APA (Table 17.2) that plagiarism is noted as violations of scientific integrity.

> Plagiarism in all of its forms is a deceptive practice and violates core values of honesty and transparency as well. The deception is pretending that one is the source of the material or idea or that the present statement in one's own work has not been provided before.

Plagiarism still seems abstract to many. To help make this more concrete, let us look at a recent study that evaluated plagiarism in graduate student writings (Vieyra, Strickland, & Timmerman, 2013). To operationalize plagiarism for purposes of the study, the authors utilized the codes listed in Table 17.5. Invoking these codes, 115 graduate research proposals were checked for plagiarism using a special software checker (SafeAssign™) that can identify plagiarism.[3] Plagiarized text was found in 28% of the proposals. The authors concluded from that the type of plagiarism is fairly common and that the frequent occurrence probably reflected the lack of awareness of the requirements of scientific writings. It is true from other work that students do not recognize the problem. For example, Internet material that is freely available occasionally is seen as not requiring citation or credit. Plagiarism generally is brought to the attention of college students early in their careers. Indeed, an entire research area has emerged that focuses on student plagiarism and strategies to combat it (e.g., Jiang, Emmerton, & McKauge, 2013; Owens & White, 2013).

Table 17.5: Example of Data Coding Scheme for Evaluating Plagiarism

Type of Plagiarism	Description
Direct Copy	Verbatim copying
Word Change	Nearly verbatim copying with a few words replaced by synonyms
Grammar Change	Whole sentence fragments copied verbatim, but writer reorganizes the order in which they appear in the sentence and/or changes verb tenses
Complex	Writer attempts to paraphrase by using multiple techniques listed above, but much of the sentence is still recognizable as copied and/or the material is not cited.

Source: Vieyra, Strickland, & Timmerman, 2013, p. 39.

> Special problems and opportunities emerge in science in the circulation of unpublished materials (e.g., manuscripts that are reviewed, convention presentations circulated in writing), and this may be exacerbated by posting materials on the Web.

PowerPoint, lecture materials, and unpublished papers or published papers with copyright can be readily found and used without providing credit. The Internet in particular provides easy access to the work of others, and cutting and pasting of passages is not that difficult. With that same technology come protections. I have mentioned that there is software that can be used to detect the likelihood of plagiarism. Many such programs are available. One journal uses software when plagiarism is suspected and also encourages authors to run their manuscript through such software before submitting a manuscript for publication (Goodman & Mallet, 2012).

Plagiarism is not a problem merely for students of course; nor is this a new problem. A Web search of incidents of plagiarism reveals a long list of best-selling authors, scientists, historians, and more with a long history (http://en.wikipedia.org/wiki/List_of_plagiarism_incidents). What is new are the ease of accessing information through the Internet, the massive amount of published material that emerges in science (more journals and scientific output than ever before), and software and increasingly sophisticated algorithms that can check on plagiarism. The solutions to plagiarism involved educating people on what that is but education, knowledge, and information alone are rarely sufficient. As a general guide, in writing always err on the side of providing full credit for any source that was used whether for idea, phrasing, or quotes. Remove all ambiguity about sources.

17.3.7: Self-Plagiarism

A less familiar variation of not providing appropriate credit is *self-plagiarism*, also noted as an issue in the ethical codes (Table 17.2).

> *Self-plagiarism refers to presentation of prior work (material, quotes, ideas) without acknowledgment and passing off the material as if it is new.*

There are many variations of self-plagiarism, including:

- Submitting the same paper a second time (duplicate publication)
- Copying select sections of text or figures and publishing those, copying from one's prior work sections
- Presenting the same data again as if it were not presented previously

> The variation that is optimally clear is presenting the complete or almost complete version of a paper or article a second time (in a second outlet) without crediting that the work was already published.

This is the scientist's equivalent of students trying to use the same term paper for two courses but without citing that the material has been presented previously. Self-plagiarism in this instance is the complete reuse of material without noting that. Another term for this in publishing is *duplicate publication* in which the same article is submitted and published to two (or more!) different outlets. In all likelihood duplicate publication is easier because there are many online journals that accept papers; many of them charge authors for publishing the article and thus have a commercial interest in accepting as many papers as possible and sometimes with little or no evaluation of the contents.

Duplicate publication is a violation of scientific integrity. It also may be illegal. Most articles in journals and textbooks require authors to sign off (sign away) copyright. That means that the owner of the work that has been printed is the commercial company or professional organization that published the textbook or journal. Citing that work—one's own work—could be a copyright infringement unless explicit permission is sought of the copyright holder. This is not usually enforced because of the low base rate of such publication, the difficulty in monitoring the tsunami of publications, and the fact that authors often acknowledge that a version of the paper or sections have been published previously.

At the other side is repetition of material that may be even close to word for word but is less clear as a problem. For example, many investigators conduct a program of research, which consists of several similar studies on the same topic often conducted over a period spanning years. In programmatic research, progress can be made as the investigator pursues topics in depth and many related areas to which they lead. As the author writes up individual studies, the methods (e.g., procedures to recruit subjects, to run them through the study, the equipment, and the measures) may be identical across studies. Understandably, the author wishes to write up the

matter in the most concise way and hence it is reasonable to "paste" in from a prior manuscript the description from a previous study. This is self-plagiarism. The way around this is not to restate the material in different words but state that the procedures have been described before and present the procedures again, perhaps in a more abbreviated form.

> The general lesson of plagiarism invariably involves acknowledging prior sources.

Self-plagiarism involving complete reuse of material or duplicate publication raises deep concern. Outside of that, self-plagiarism is not at the same level of concern as stealing or ideas from the works of others without giving credit. The concept of stealing from oneself or reusing material is interesting when one moves outside of science. In music, dance, and painting, for example, it is often the case that great composers, choreographers, and painters take material from one or more of their prior works and insert it, changed but sometimes not changed, into a new piece (e.g., Baserga, 2011). If one knows really well the music, choreography, and painters of a specific artist (e.g., Wolfgang Amadeus Mozart, George Balanchine, Pablo Picasso), it is not difficult to identify these recurrent themes and multiple concrete instances of reusing material, passages, and so on. Self-plagiarism is not raised in the arts very much.

In science writing or communication (presentations), authors more routinely note (in a footnote) that material provided in this paper (chapter, textbook, article) has been partially presented before. Such an acknowledgment is the default position. When in doubt, give credit. Acknowledging reuse of one's own prior material can be easy. Your footnote can say something like, "I am not just brilliant in this paper—I was brilliant in the same way in another paper a few years ago when these ideas were first stated" and then cite the paper and include that paper in the References. If you do not like my wording, still cite the paper explicitly or as you read a study see how the investigator communicated that something in the present study appeared in a prior paper.

17.4: Authorship and Allocation of Credit

17.4 Recall scientific ethical codes as applicable to authorship and allocation of credit

Plagiarism clearly is an instance of not allocating credit where credit is due. There is another manifestation where allocation is an issue that is quite separate. A critical issue in the allocation of credit pertains to the credit accorded to colleagues and students involved in research. Projects are usually collaborative and collaboration can vary in scale. Some projects are collaborative in the sense of involving multiple investigators, from multiple sites and often now different countries. In addition, studies often require specialist (e.g., running a neuroimaging center and programming the imaging software, genetic analyses, help in conceptualizing the study or developing the computational or math model), and ensuring its completion. It is not rare to have some papers (e.g., genome) with more than 20 or 30 authors (e.g., as in early gene sequencing studies).[4] At the other extreme, an individual investigator conducting research in her lab has assistants, students, and postdoctoral researchers, and they too are collaborating to bring a given project to fruition. For psychology, more commonly are collaborative studies involving a handful of individuals working on a given study. These may include a mix of individuals with different levels of training (e.g., senior investigator, postdocs, lab assistants) and varied roles. That diversity in training and roles adds a few complexities to the topic of allocating credit.

> *Allocation of credit emerges in deciding whom to list as authors on a research article, the order in which they are to appear, the relation between junior and senior scientists or faculty and students, and how the different roles and contributions affect authorship.*

I am discussing this issue in the context of scientific integrity but it overlaps with personal integrity, i.e., the character or nature of the individuals involved and going back again to the Einstein quote at the beginning of the chapter.

We begin with the most salient issue in allocation of credit in this context is whether a person involved in the project in some way ought to be listed as one of the coauthors on any publication and report.

> Decisions about allocating credit and authorship are fraught with human frailties related to status, power, greed, ambition, insecurities, anger and revenge, and personality style of the investigators, collaborators, and research assistants.

Surveys reveal that a frequent source of disputes among authors and concerns relate to not receiving appropriate credit for contributions (Benos et al., 2005; Seeman & House, 2010). For example, one survey found that 50% of respondents believed they did not receive appropriate credit for their contributions to the published report (Seeman & House, 2010). There is another side to all of this that is less visible. Many individuals surveyed reported not knowing that they were an author on the paper until after the paper had been submitted. Also, many individuals finding out they were going to be an author ask to have their names removed. These comments are designed to make a few salient points. The processes of authorship and contributing to a project have not been very clear or transparent.

And, there is no standard rule or guideline that is invoked or authority to whom one can appeal.

The human frailties often are heightened in relation to authorship issues in large part because the stakes of authorship are high or perceived as high.

> Publication and the number of publications can have direct implications for faculty tenure, promotion, and salary at many universities.

Being an author on another published article moves one closer to these rewards, and hence few researchers early in their academic career would say, "It does not matter whether I am an author, I can be on the next one or the one after that." The stakes have moved down developmental stages so that authorship issues can be important even earlier in one's "career." Undergraduates who apply to graduate school will have their application materially enhanced by being on a publication or two. And surprisingly many undergraduates do. Some graduate schools in psychology often want to know the equivalent of whether the applicant can do research, likes research, or has any understanding of or experience with the research process. A couple of publications take that issue off the table. Graduate students applying for postdoctoral positions or faculty positions too can be helped by publications and so on with promotions once on the faculty. One hopes that "having publications" does not move down and up the developmental spectrum so that getting your son or daughter into a good preschool or getting your grandparents into an assisted-living facility will be helped if they had a publication or two. Scary.

Apart from the consequences associated with authorship, there is a broader issue, namely, the just and fair allocation and recognition of credit and that can be evaluated independently of other issues. Credit ought to be given where credit is due apart from the professional publishing game. The difficulty is that once one moves beyond stating such generalities and descends from the high moral ground, chaos can prevail. Individual differences, temperament, personal insecurities, and other motives noted previously color all of our interpretations about what is a contribution to the study and how much that contribution ought to count toward the final allocation of credit. It is easy with the limits of one's perception to feel (if not say), "I eagerly give credit where credit is due but you (dear student, colleague) do not qualify."

17.4.1: Guidelines and Best Practices for Allocating Authorship

There are many guidelines for allocating authorship, and these are fostered by different journals and professional organizations. A prominent set of guidelines is provided here because of the large following. The International Committee of Medical Journal Editors (ICMJE) is a large group of individuals who represent journals and organizations in biomedical research (e.g., medicine, health). The group meets regularly and considers critical issues and makes recommendations in conducting, reporting, editing, and publishing scholarly work including issues related to scientific integrity (ICMJE, 1997, 2013). The overall goal is to identify best practices for conduct, reporting, and ethical standards for research and provide a valuable reading experience (please see Further Reading).

For purpose of this discussion, it is useful to note the recommendations that specifically focus on the conditions or criteria for being an author on a paper.

> Four criteria are noted. Anyone designated as an author ought to have made substantial contributions:
>
> - To the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work
> - In drafting the work or revising it critically for important intellectual content
> - In providing final approval of the version to be published
> - In addition, the person must agree to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved (ICMJE, 2013)

The recommendations are clear in noting that all coauthors should meet all four criteria and that anyone who meets the criteria ought to be coauthors. The first three criteria focus on the scope of work or effort. The last criterion is designed to ensure that people do not just have their names placed on the article without knowing or without have responsibility for the contents.

Many individuals may work on a study but do not meet the criteria. The ICMJE guidelines note that those who have not met these criteria but contributed in some way might warrant designation in a footnote. Such tasks as entering data, running the data analyses under direct supervision, and preparing ancillary materials (e.g., references) might warrant acknowledgment in a footnote, but are not likely to be considered a professional contribution in the sense of the other criteria noted previously.

Although the focus of the organization is on biomedical research, the ICMJE recommendations and guidelines on publishing are widely applied and endorsed beyond biology, health, and medicine. Hundreds of journals have subscribed to the recommendations and many journals in psychology, psychiatry, and psychotherapy among their ranks (for a full journal list see www.icmje.org/journals.html). The recommendations are not enforced or policed. Yet, at the point that a manuscript is submitted for possible publication, many journals require the authors to specify precisely what role

each has played in the study and preparation of the manuscript for publication. In keeping with the fourth recommendation, each author may be asked to sign off to attest to his or her role and approval of the study.

Recommendations and guidelines as I noted before are not mandates or rules. An extensive review of authorship practices, contributions, and dissatisfaction has revealed that many of the recommendations are followed, but coauthors also voice dissatisfaction and departures from the recommendations, on many occasions in which they feel they have been unfairly treated (Marušić, Bošnjak, & Jerončić, 2011). The recommendations cannot be expected to address quite different perceptions that individuals have. Any guideline is likely to be subject to interpretation and, for example, what constitutes a substantial contribution to a study can differ greatly in the views of those who helped bring the study to fruition. A major contribution might stem from meeting one criterion extremely well. Or if a potential author meets two or three but not four of the criteria, she is still likely to believe strongly that authorship is warranted and might well be.

I hasten to add an equivalent issue emerges in allocation of credit and giving out Nobel prizes. Often a given prize is shared (with a limit of up to three people, all of whom must be living). It is often the case that when one person receives the award, some view this as an omission of a second equally deserving person; when two people receive the award, some view this as an oversight of the third person or even that the third and fourth persons really should have been given the award instead. (Also, if one had died, one cannot receive a Nobel prize—seems unfair—most of the time dying is not the person's fault and should not diminish his or her contribution. An exception is if the Nobel prize has been officially announced and the person is living but dies before actually being handed the prize.) Allocation of credit requires judgment and consensus in giving out Nobel prizes. Allocation of credit in a given lab with one or two senior investigators is more top down and authoritative and does not require the input of many people.

Whether a particular criterion for authorship is met is subject to interpretation. For example, if someone analyzes the data and performed exactly the analyses asked by the investigator, then:

- Does that warrant authorship, footnote, or no credit?
- What if data person noticed something in the analyses?
- Did some other analyses, and raised a novel issue that finds its way into the manuscript?
- What if another person wrote up the Method section and the References?

Now add to that they also drafted the Introduction too with background theory and research. Minor clerical tasks versus huge conceptual contributions are clearly perhaps footnote and authorship worthy, respectively, but the huge area between is gray that leads to quite different views about what is fair.

Let us say that one is going to be a coauthor. Other potential sources of frustration can await us. The authors on the paper must be listed in some order—rarely is it alphabetical. Rather, someone is judging who should be listed as a coauthor and in what order. Perhaps this ordering is based on the extent of the contribution, clearly another judgment call. The guidelines do not help on ordering of authorship—someone is judging the value, amount, and quality of a contribution and that amounts to discretion of the senior investigator. Coauthors may feel that credit has not been allocated fairly as reflected in the order in which authors appear on the article. Their indignation often is well placed, as evident in honorary authors, as discussed in a moment. One can only advise contributors to make it as much as possible explicit at the beginning of a project.

The recommendations and the requirements serve a broad purpose by making more salient that there are criteria to serve as guidelines for authorship, when collaborators in research have a reasonable basis to expect authorship, and that being included or not included in the list of authors is not purely fiat by the lead or senior investigator. Table 17.6 provides a useful guide to the discussion of authorship for a given project. The table makes explicit what parts of authorship discussion should be taken up and when. Not discussing these topics explicitly creates great risk of disappointment, aggravation, and ill will. The person most likely to suffer is the junior person (student, younger investigator) on the project.

Table 17.6: Best Practice Principles and Recommendations for a Fair Allocation of Authorship Credit

Convenient time	Recommendation
Before Research Starts	**a.** Decide who will be the author(s) **b.** Define responsibilities of the authors **c.** Ask technicians, statisticians, software developer, and other individuals involved, whether they are interested in authorship
	Principle
After Manuscript Preparation, Before Submission	**a.** Create contributors' list and determine relative contribution **b.** Determine authors in the byline list **c.** Determine guarantor* **d.** Determine corresponding author **e.** Disclose contributions

*Guarantor is the person who assumes responsibility for the integrity of the work as a whole from design through publication.

Guidelines ought to be complemented by commonsense seasoned with wisdom. From my perspective, consider collaboration in research as a relationship—this is not warm fuzzy clinical platitude but has a broader point. Collaboration, when cast in a relationship light, prompts useful concerns. Be careful with whom you are involved. As in a relationship, you might like to know where this person has been before and with whom (other collaborators and their views) and seek personal qualities that serve as a solid base. If you can work with gracious, non-pontificating, open, and generous investigators, leap to the opportunities. And if you find yourself in a relationship with the opposite type, perhaps leap in a different direction. It would be helpful if you are working with someone who is easily approached and even fosters the discussion (of material in Table 17.6). We often need to select advisors for other reasons (e.g., in my case finding someone with expertise and research projects on mindfulness among bipolar zebra fish), but after completing one's degree requirements it is easier to select to include generosity of spirit, fairness, and more into the criteria for selecting collaborators.

> Investigators, whether faculty and students or multiple colleagues within the profession, are encouraged to discuss these matters explicitly at the inception of research, not merely to address issues of authorship, but also to decide tasks to be completed, responsibility, and credit in relation to all facets of the study.

Explicitness is important, but perhaps assumes increasing importance to the extent that there are power, age, status, or other such differences among the investigators and collaborators. The greater the differences, the more assumptions are likely to be made and the less likely such assumptions are to be checked with direct communication. Converse with colleagues, advisors, and mentors if you have any interest in subsequent credit. Make your concerns about authorship clear and explicit as early as possible and your own style is relevant to how well this is done and how clear the results are.

In some collaborative arrangements, it is difficult to approach a senior collaborator or advisor. Consequently, it is useful to place responsibility on senior investigators for:

- Initiating the topic of allocation of credit
- Addressing the matter directly
- Encouraging dialogue on the topic

Of course, we know from experience or understanding of human nature that those senior investigators who take the responsibility and begin with open communication about such matters are the collaborators we may need to be least concerned about in relation to allocation of credit.

17.4.2: Special Circumstances and Challenges

A few special circumstances in authorship are worth noting in passing. They raise concerns about scientific integrity because in various ways they may misrepresent a critical facet of the research project and the contributions of investigators. The first circumstance is referred to as *honorary or gift authorship*. These terms are used for individuals who are added to the list of authors but who have not contributed to the conception and design of the study, the collection, analysis, interpretation of the data, and drafting of the article (Bates, Anić, Marušić, & Marušić, 2004; Smith & Williams-Jones, 2012). That is, consider these individuals to have met none of the authorship criteria noted previously.

As examples, I have seen spouses and partners (who are in the same or different department) and administrative bosses (department chairs, senior hospital administrators) added to the list of authors on a manuscript. They seem to be authors ex officio, i.e., by the rights of their position. Sometimes, senior investigators who facilitated a study (e.g., by allowing access to a patient sample) might be placed on the publication as a courtesy. This is rare to nonexistent in an academic psychology department and more likely in an academic medical school department or hospital where providing access in fact may be considered a huge contribution. Although comparative data are not available across various sciences, a survey of six general medical journals revealed that approximately 18% of the articles included honorary authors (Wislar, Flanagin, Fontanarosa, & DeAngelis, 2011).

> Honorary authorship is viewed increasingly as inappropriate in light of the criteria that are explicit about what justifies authorship, as discussed previously.

Also, when a manuscript is submitted and the role of each author in the project has to be described, either the honorary author has to be omitted or her or his actual contribution needs to be inflated (misrepresented) to justify inclusion on the paper. The ethical issue is nontransparency and deception over who contributed what to the study. Consequently, merely slapping on someone's name who did not contribute is a violation of scientific integrity.

There are hazy areas for sure. Say I have collected all of the data, written up and published several studies, and now you come along and use the data in a different way. I do not contribute to your study or write up in any direct way but none of the work could have been done without me. In fact, I obtained grants to completely support obtaining the data long before you arrived on the scene. What is appropriate credit? One related recommendation now is to include the original investigator as an author if data are reused or used in novel ways for a new study (see Cooper &

VandenBos, 2013). The reason this is hazy is that in a hospital setting an administrator or director might appropriately feel the same way, namely, this research could not be completed without me. I did nothing directly to contribute to the details of the study, but I allowed access to the population and setting. I might consider my contribution by virtue of being the administrator in charge of this program. As we start to quibble about authorship, I may gently remind you, "try doing a study without subjects." I am not advocating honorary authorship—just the opposite. But one should be aware of the different sides and perceptions.

> Another authorship issue that raises ethical issues is referred to as *ghost authorship*. This refers to someone writing up a study in whole or in part but is not named as an author or noted in the acknowledgment section (Ngai, Gold, Gill, & Rochon, 2005).

A survey mentioned previously found that approximately 8% of the articles from major medical journals included a ghost author (Wislar et al., 2011). Obviously, outside of surveys, it is difficult to identify ghost authors by definition—they do not list themselves, dress up as ghosts on Halloween, and in fact do all they can to be stealth. (As a momentary break for irony, in this section on Allocation of Credit, we can see to odd scenarios—some people [coauthors] desperately plead to get the credit they deserve and be listed as an author and view being on a publication as a path to a better future. Now switch to ghost authors who actually write up the study, sometimes completely, and desperately try to avoid any public recognition or credit.)

The usual context for ghost authorship is industry where results of clinical trials for a procedure or medication are being written up. Before or after the write-up, an expert or two are hired, often paid handsomely, and serve as "authors" for the paper, whether they have had a role in the study or not. That is, the authors listed on the manuscript are paid consultants who may not have played any role in the study or write-up. The payment is for permission to use their names as recognized authorities in a given area of work. Court documents have identified the practice and its relatively common use (Ross, Hill, Egilman, & Krumholz, 2008). Ghost authorship raises multiple scientific integrity issues, and in some countries (e.g., Denmark) it is formally designated as scientific misconduct. The practice is:

- Deceptive
- Violates transparency and honesty
- Gives no accountability of the "authors" for the study and its procedures

Most salient is that ghost writing usually reflects a conflict of interest. Industry ghost writers have a position with regard to a procedure (e.g., benefits of a new medication, or medical device) and may not be objective in portraying the results. Ghost writing can influence what material is presented and emphasized and the "spin" on the findings. We already know that research sponsored by industry is more likely than nonindustry-funded research to show the beneficial effects of a medication and to show that allegedly harmful food products (e.g., sugary soft drinks) are not that bad (e.g., Bourgeois, Murthy, & Mandl, 2010; Lesser, Ebbeling, Goozner, Wypij, & Ludwig, 2007; Vartanian, Schwartz, & Brownell, 2007). Negative or mixed results about a product can actually harm a company (e.g., strong effects on the stock value of a company) (Rothenstein, Tomlinson, Tannock, & Detsky, 2011). Ghost authors write with complete awareness of the implications, financial rather than scientific, and cast the findings in the best and possibly biased light.

Ghost authorship has various solutions:

- To prohibit the use of ghost authors
- To ensure that it is clear in the published report who wrote the article in whole and in part

This can be accomplished by including the authors (so there is no ghost) or including that person in the acknowledgment with a clear statement of what role the person played. As I mentioned in comments on honorary authors, journals more uniformly ask when a manuscript was submitted what the precise role of each author was and then agreeing to be held accountable for the material included in the article. This practice is intended to decrease honorary and ghost authors as well as surprised authors, i.e., those who were listed on the study but had no idea they were to be listed.

17.5: Sharing of Materials and Data

17.5 Review the advantages and disadvantages of sharing of materials and data of scientific work

A core value mentioned previously included transparency of scientific work. Among the many meanings of this is openness about what one has done, how, and with what tools. This is translated into sharing of materials and data with one's peers. This is not sharing or show and tell for its own sake.

> *A central feature of science is the ability to replicate the work of others.*

Replication is usually discussed in the context of repeating the procedures of a prior investigation. In relation to scientific integrity, there is more to it than that. Replication begins with the obligations of an investigator to provide colleagues with the materials to permit them to

conduct replications. This might entail providing further descriptions of procedures and specific measures, responding to various questions that facilitate replication of the study, or making available materials used to score, code, or analyze the data. As the ethical code of APA notes (Table 17.2), sharing of materials and data is the explicit responsibility of the investigator.

Often critical features of a study have required years to develop (e.g., treatment manuals) or have important financial implications (e.g., proprietary software, a new psychological test) that make investigators reluctant or occasionally unwilling to share materials. However, the obligation to share materials begins when the individual enters the role of scientific investigator and places his or her work in the scientific domain (e.g., presentation of a paper, publication of a scientific article). At that point, the investigator has entered an implied contract with the rest of the scientific community in which he or she will aid in continuation and evaluation of the research. As a reader or consumer of a particular article or study, it is quite fair to write and ask for materials or further information.

One of the most frequently discussed issues pertains to the sharing of data. This, too, is related to replication. Can one obtain the same or similar findings when analyzing the original data that were published? A colleague may believe that the original data were not analyzed correctly or in the most appropriate fashion, or would lead to quite different conclusions if analyzed differently. Data are viewed as part of the public domain and to be shared with others if requested.

There is often reluctance of investigators to share data. One reason is that a given study may be drawn from a larger database. Several other projects may be planned, and the investigator may be unwilling to circulate the data until the projects have been completed, analyzed, and reported. Here too once an article has been published in a scientific journal, it is difficult to justify withholding the specific data set on which that article was based. That data set, even if not the entire database, might well be considered to be part of the information available to the scientific community.

> For federally funded research, it has become law to make raw data available (referred to as the Shelby amendment and passed by the U.S. Congress in 1998). The law requires that the public (anyone who asks) be given access to data generated by federally funded research.

Concern grew from instances in which the government was refused access to data, particularly in relation to controversial policies (e.g., clean air standards set by the Environmental Protection Agency). Parties that might have keen interest in the policy and its rationale (e.g., business and industry that might want to challenge the standards) ought to have access to the information (data) on which conclusions were drawn. That is, if government is to make new rulings that could affect business (e.g., pollution standards), the businesses ought to have access to the information on which the policy is based. We know as scientists, quite apart from any policy, that the same data set might be subject to different interpretations and indeed if analyzed differently might support different or new conclusions. Access to the data allows all parties to see what was done and to make judgments about the conclusions.

At first glance, the issue may seem to be obvious—of course all data should be shared and if taxpayer money (federally funded projects) is involved, perhaps there is a special right to public access to these data. There are pros and cons about sharing data despite agreement on transparency, honesty, and access to materials (Koslow, 2000; Piwowar, Day, & Fridsma, 2007).

> The sharing advantages are in keeping with the values of science. What is the other side?

Among the issues are concerns that:

- Use of the data may violate consent of the subjects who did not agree to new or additional studies
- Investigators who conducted the original research may not be able to write up all of their studies before others who receive the data set early write up those originally planned studies
- Interfering with the study (if the results are in the public domain while subjects in a longitudinal study are still being run) and if the early findings could bias later findings or alter participation in the study (e.g., attrition)
- Providing access to proprietary materials or inventions that may harm in some way (e.g., credit, financially) those who developed the materials
- Sharing data and materials often has costs and not merely a matter of sending an electronic file, and who bears the price of those costs and genuine dangers associated with sharing data and information (to be addressed below in the discussion of dual use)

The default position is to share one's data and information to qualified professionals who make the request. Also recommended is that in sharing the data, authors agree formally (in writing) in advance of sharing what will be done, how the information will used, and who receives credit (APA, 2010a). As I mentioned, publically funded research requires (federally mandates) that the materials and data are available to those who are interested. For example, the National Institute of Mental Health (NIMH) provides access to data sets from clinical trials, including of course trials on topics within clinical psychology (e.g., covering a variety of psychiatric disorders and interventions for children and adults) (see www.nimh.nih.gov/health/trials/datasets/index.shtml). Access to the data is

provided to qualified professionals who must complete a formal request and agreement to access the data. Most psychological research is not funded by grants and does not have that explicit requirement.

Second, journals occasionally require that any article accepted for publication make available the data on which the article is based. The data set on which the study was conducted may need to be submitted along with the manuscript accepted for publication. It is now possible and feasible to collect, store, and make readily available data in light of electronic files and online storage services.

Third, increasingly in research, databases are made publicly available to other researchers to permit further analyses (see Stewart, 2012). Typically, these are from large-scale studies where extensive information is collected. The researchers responsible for the studies may have a set of studies in mind for which the data were collected. Yet, the database is made available with the expressed goal of utilizing the rich data set and drawing on the creativity of many investigators to extract further information. For example, the National Comorbidity Survey (and Replication; NCS and NCS-R, respectively) are large-scale evaluations of psychopathology, course, risk and protective factors that has generated scores of publications that have greatly added to our knowledge about disorders over the course of life among adolescents and adults (see www.hcp.med.harvard.edu/ncs/; National Institutes of Health [NIH], 2013g). Not only data sets but also measures used in the study are publicly available and research can access the information for use.

Related, the NIMH Collaborative Psychiatric Epidemiology Surveys makes available data on the distribution, risk factor, and other characteristics with emphasis on minority groups (www.icpsr.umich.edu/icpsrweb/CPES/). This project brings together three national databases: the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS). The project permits the in-depth study of cultural and ethnic influences on mental health. Data, code textbooks, and descriptions of procedures are available for use.

Often groups of researchers with similar interests agree to share data and to combine their data. These consortia are not formal mechanisms but rather emerge from collaborations, shared interest, and realization that much better progress can be made on a topic if data sets are shared. Consortia do not spontaneously emerge very often, but the benefits are well recognized. To that end, grant agencies occasionally provide funds to build these sharing networks (e.g., The Human Connectome Project, Biomedical Informatics Research Network, and Alzheimer's Disease Neuroimaging Initiative) (Gorgolewski, Margulies, & Milham, 2013; Mennes, Biswal, Castellanos, & Milham, 2013). There is renewed interest in combining or pooling raw data because of the advances in the ability to analyze such data sets and in the software for combining and evaluating the data. Pooling data, as it is sometimes called, from many studies changes the scale of science (Hussong, Curran, & Bauer, 2013).

In short, there are several mechanisms in place for data sharing. Funding agencies, individual journals, and open access to databases are influences that foster or actually require sharing. These add to the APA professional codes that make sharing of data and materials explicit as a responsibility of an investigator. Yet even with all this in place, investigators often do not receive the data or materials they request from authors (Campbell et al., 2002; Rathi et al., 2012). Also, journals vary greatly on the extent to which they require data sharing and monitor or enforce the policy when they do have a policy, so some of the mechanisms in place are not invariably effective (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011). Investigators give as reasons for not providing such materials as follows:

- Providing the materials requires too much cost or effort
- Protecting trainees so that they could publish additional studies
- Protecting their own right to publish

The reluctance to share is in fact related to the strength of the findings of the study. Reluctance to share is associated with weaker evidence and a higher rate of errors in the study (Wicherts, Bakker, Molenaar, 2011). Thus, one conclusion from this is that investigators are reluctant to share because they are concerned that reanalysis will generate different conclusions or expose errors that were made that change the findings. Clearly challenges remain to foster universal and routine sharing of data, but I believe there is clear movement toward that goal.

17.5.1: "Big Data:" Special Circumstances Data Sharing

Data sharing usually means I give you access to my data so that you can scrutinize my analyses, conduct new analyses, or generate new studies. There are new and special circumstances that raise novel issues. Currently in science, health care, and business and industry, there is great attention to "Big Data."

> *Big data refers the harnessing of massive amounts of information that is available and utilizing that in novel ways*. Big data in part relies on the fact that enormous amounts of data are being collected in public life.

Another equally critical component is that massive of amounts of data can be shared, used in networks accessed by many, and stored (e.g., in the "cloud").

Perhaps the immediately familiar use and collection of big data is in business and industry. Information about our purchases and the sites we visit on the Web provide data to many other sites we do not know about that build a profile of:

- Who we are?
- What we like?
- Characteristics of our personality
- Where we are within a very narrow geographical range?

The data are combined and accumulated, and a full picture of millions of people can be obtained. This information is used to target advertisements and opportunities. So in the simplest version of this, say we bought organic coffee online. Now we can look forward to ads for coffee makers but many other unrelated things (e.g., small appliances, scores of health products or foods, vitamins, natural skin conditioners, and so on partially triggered by our search for "organic" coffee). The "unrelated things" might be just our perception because computer algorithms (and assorted correlations) might establish that people like us or people who search for and purchase product *x* are more likely than the general population to also purchase *y* and *z* products. More appliance and food ads come to us online and maybe even coupons arrive in the mail from nearby stores that sell small kitchen appliances or organic foods. This is barely the scope of big data and its uses. Indeed, we have learned that many nations are monitoring the phone calls and communications of leaders of other nations and on a scale that involves millions and millions of such calls. Spying now involves big data.

The sources of big data include the many traces we leave primarily from our use of smartphones, tablets, and computers to browse or use social media (e.g., Google, Twitter, and Facebook) or look up information. Data are available on our:

- Movements (Global Positioning System)
- Interests
- Credit card purchases
- e-mails
- Social behavior
- Blog postings
- Blog visits

Big data also includes combining multiple data sets. In addition to the other sources of data I mentioned, we also have academic, and physical and mental health records, and more. Massive data sets can be combined to produce unimaginable super massive data sets. Clearly the availability of data is not just more of the same data with a larger sample size we have been collecting but has special features such scale of participants, variables that are studied, and ability to process massive amounts of information.

To be more concrete, consider a relatively recent example of the collection and use of big data. The outbreak of flu in 2009 and its spread in the United States was greatly facilitated by Google's use of big data (Mayer-Schonberger & Cukier, 2013). The idea was that as people searched the Web, the specific words and phrases they searched would relate to (correlate with) their contracting flu symptoms. Google identified terms that people were searching for by focusing on flu symptoms words and found that search terms were a good indicator of flu activity (see www.google.org/flutrends/us/#US). Search terms combined with information from the Centers for Disease Control and Prevention (CDC) and the addition of sophisticated math modeling were able to identify the distribution of the flu.

How is this "big data?"

Google was able to look at over 50 million search terms people use. The task was to use computer power to identify spread of the flu and people's search terms associated with that and then use the information to identify in real time where and when the flu had spread. Google processed over 450 math models to compare their predictions of search terms with actual flu cases in previous years based on CDC information. Thus, one could test the relation of search terms with the known spread of a prior flu. Eventually 45 search terms were identified and entered into a model that could tell where the new flu outbreaks had occurred and were occurring and in real time. The process trumped the old way (doctors and hospitals reporting cases weekly for example) to monitor the problem and more importantly to intervene where needed.

The brief example is useful in making more concrete that "big data" does not merely refer to larger sample sizes and more measures than we usually have used in studies. Also, one can see that big data goes beyond making research databases available to other scientists, as discussed previously. The scale of these measures and tracking continuously and the data-analytic tasks and challenges make this a qualitative change in how science is conducted (National Research Council, 2013).[5]

Big data raises new hopes to:

- Address health issues (e.g., causes of disease and many different paths to a particular disease)
- Genetic and epigenetic variation, and its impact on all facets of life

For example, the spread of disease in developing countries has been greatly enhanced by identifying when people come and go and where they are based on the location of their cell phones (e.g., Buckee Wesolowski, Eagle, Hansen, & Snow, 2013; Wesolowski et al., 2012). Similarly, cell phone data were used to evaluate evacuations after a massive earthquake in Haiti (Bengtsson, Lu,

Thorson, Garfield, & von Schreeb, 2011). The data provided information on the spread of cholera too as people's whereabouts and comings and goings could be tracked.

17.5.2: More Information on "Big Data"

> In relation to psychology and clinical psychology, there are many obvious uses of big data that have yet to be fully exploited. Perhaps the most obvious will be in neuroscience.

Social, cognitive, cultural, and clinical psychology rely increasingly on identifying underlying neurological processes of conformity, decision making, and various psychological disorders, and cultural variations of these. Neuroimaging is advancing enormously to make it so that there is whole brain imaging and imaging networks and connectivity at a given point in time and with dynamic changes over time (see Turke-Brown, 2013). The multiple views of the brain in these ways will generate massive amounts of information about networks and changes of diverse combinations of neuronal activity. (Remember there are about 86 billion neurons in the brain and the networks and their interactions remain to be worked out. The range of possible combinations and their interactions in changing states is enormous.) As the level of analyses involves these combinations of neurons and networks and at different levels (e.g., molecular, system), the amount of data will soar. Complex software, math, and statistics to handle the big data will be central to advances and all of this applies to various facets of psychology. Also important to note is that big data are not all the result from data in the public domain and as a result of various records (medical, criminal, social media).

Big data, whatever its source, moves beyond doing the usual at a different scale. Domains not usually discussed in one area (e.g., climate, weather, temperature, humidity, chemical changes in the air) are among many domains that could relate to mental and physical health, interpersonal relations, and violence. Actually, we already know that they do in human and nonhuman animal studies (e.g., Bolton et al., 2013; Guxens et al., 2012). Yet, big data will go beyond an isolated study on the topic but can do massive exploratory analyses to identify relations, control variables, and develop and test models that provide insights not otherwise available and then add other domains (e.g., genetic influences) to look at moderators and mechanisms.

I mention big data but because it gives a greatly expanded meaning of sharing of data. Also, big data raises a whole new set of ethical issues. Massive information is being collected that the public does not even know about, leaving aside providing consent. Also, many already big databases will be combined. I mentioned in the previous chapter an illustration where a country is combining genetic data (from DNA) and health records of the population with the goal of better understanding disease. Already important insights were generated in published studies, but the project came under further scrutiny. Participants (citizens) who provided the genetic data did not consent to the broad use of combining the information with other data. The matter went to the courts that (at the time I write this) ruled against sharing without consent (Kaiser, 2013a).

Big data does not mean that the identity of the individual is lost. Indeed, the use of big data in homeland security and antiterrorism is not designed to look at "trends" or "intriguing findings" but rather to identify people and groups very specifically. The ethical issues for the public include informed consent, invasion of privacy, and deception (use of information in ways never disclosed). Elaborated ethical codes will be needed to balance uses of big data and privacy and protection of individuals and groups of participants. Universities also have and use big data—e-mail and Internet use that go through the university servers—all tracked and can be identified if there is a cause (e.g., public health and safety). Universities often are quiet about all of this but tracking without formal consent is in place.

17.5.3: When Not to Share Data

> Transparency, data sharing, and being open all sound wonderful in conducting one's research and are part of the core values. There are exceptions that have received increased attention and warrant comment because they argue against data sharing or for restricted, cautious sharing. These exceptions have emerged outside of clinical psychological research. Yet they are important to note because they convey the need to ponder multiple issues of one's research.

Also, any ethical codes or guidelines and regulations related to these exceptions are likely to apply to psychological research. Added to that, once we cover the issue it is easy to envision psychological research where sharing of information might purposely be restricted.

First, scientific studies, often funded by the federal government, focus on very sensitive topics that relate to intelligence, technology used by military, and presumably other topics considered to be central to national security. Presumably topics such as chemical warfare and protections, novel techniques to monitor military communications, and anti-anti (add a few more anti's) jamming or missile defenses are among such topics. The research and its uses directly focus on topics in which the security and secrecy are considered to be essential and where the success of any finding, invention, or technology will be compromised if the information became widely known. There

are journals in which at least some of this research is published (Malakoff, 2013). These journals are not in the public domain, and access is restricted to individuals with special clearance.

An example that does not have full secrecy but conveys the issue is the *Journal of Sensitive Cyber Research and Engineering* (http://cybersecurity.nitrd.gov/jscore). This journal seeks to balance the natural tension between the sharing of information among scientists and protecting the circulation of sensitive information that we would not let others (enemies and unreliable friends) to know about. But other journals are not in the public domain because the information they provide is for very restricted use. There is not much more to say because information is not readily available about them. It is important to note here that there are instances in which scientific data and information purposely are not shared or at least not available to all scientists to replicate, evaluate, build upon, and so on.

> Also, it is readily conceivable that psychological research related to identification of terrorists or spies and the treatment of prisoners might inform critical issues that also fall within national security and hence are not shared.

Second, the notion of dual-use research has become prominent in the past few years as a different but related class of studies where information may be restricted.

> *Dual use refers to research that provides knowledge, products, or technology that could be directly misapplied by others and could pose a threat to public health and safety, agricultural crops and plants, animals and the environment, or national security* (NIH, 2013e).

Stated another way, the dual-use dilemma, as it is sometimes called, refers to circumstances, "in which the same technologies can be used legitimately for human betterment and misused for bioterrorism" (National Research Council, 2004, p. 2). The dual-use case might be distinguished from the security and military concern raised previously. In the security and military example, the goal of the research was directly to address secretive issues to protect (or presumably harm) and might be called single (or uni-use) research. In dual-use research, a scientific finding clearly could be used for positive or negative public good, but the goal was not necessarily related to security issues. (This discussion omits mention of my dissertation that has been referred to as the first "zero-use" research.)

The issue of dual use received major attention with recent research on influenza (H5N1 virus) conducted separately by two teams of researchers (see Berns et al., 2012; Fouchier, Herfst, & Osterhaus, 2012). The virus usually infects birds; the researchers showed how this could be modified so that it infects mammals. The scientific advance pertains to understanding virus mutation and transmission and has implications for prevention, treatment, and risk of pandemics. The scientific merit was not challenged and not the key issue here. Indeed, arguably the science reflects a critical breakthrough in understanding viruses.

> The dual-use issue was the prospect that if the methods and procedures were made explicit in the usual way they are in science that information could be used to make a bioweapon, i.e., purposely developing and circulating a new virus as part of a public, terrorist attack.

That is, one might construct a virus that could infect large swaths of the population and one for which there was no effective treatment (because it had not been in mammals before). The fear was fueled by prior anthrax (a lethal bacterial disease) scares in the early 2000s when anthrax spores were sent through the mail to many people (e.g., politicians and others) as part of a bioweapon attack. These spores when inhaled or ingested multiply quickly and can cause the disease and death.

The influenza virus findings caused international upheaval in the sciences because of reservations to publish solid scientific findings and set into place multiple efforts to develop guidelines for publication and treatment of this type of research more generally. In the United States, government agencies that fund research (e.g., National Institutes of Health) and independent panels (e.g., National Academy of Sciences) began to consider the issue and to provide specific guidelines regarding when research has to undergo another level of review specifically to address dual-use potential (see Malakoff & Enserink, 2013). The threat of dual use requires considering the implications for the findings and the procedures that were used to obtain them. The matter is still an active topic of discussion and consideration. The nuances are not only the balance of sharing information in science versus security for any single study. Presumably a single study or a line of work the study promotes might have extensions or next steps that are likely (or not) to be dual-use findings.

Misuse of scientific information is the dual-use dilemma and has its variants in different disciplines (Malakoff, 2013). In each case, the dilemma—will sharing information as dictated by the core values and practices of science lead to consequences that outweigh that practice? For example, in archeology there is concern that posting findings of newly discovered sites may be used by poachers and looters. When information about a new site has been posted on a Web site, thieves have been able to connect the photos to other mapping images, discover the exact location, and then follow up by looting the site. In conservation biology, there is reticence in sharing data about where a new or rare species has been found for fear that hunters or collectors will pursue the species (Malakoff, 2013). Technology advances have similar issues of dual use, such as:

- 3D printing can create customized objects designed by digital instructions.

- The copiers can be used to make 3D objects by building them layer by layer.
- The materials can include plastics and metals and models of neurons, living cells, and organs (not musical ones), and more.
- There are enormous real and potential uses of printers in manufacturing, engineering, and medicine (e.g., making machine prototypes, building models of human organs).
- The printers can and have been used to make plastic guns, and these guns cannot be detected by the usual security procedures (e.g., at airports) because they have no metal parts.

Perhaps these printers might be called dual-use machines. 3D printers are not readily available as are the more familiar printers we use, but that is anticipated to change. New security methods will be needed to try to control misuse of such printing, but those methods and guidelines cannot easily keep up with technological advances (*Science*, 2013).

I mention these examples only to convey that transparency, sharing data, and openness to other scientists and the public are under increased scrutiny in part by novel findings and technologies but also the ability to circulate and obtain information more widely than ever before. The circumstances I have outlined illustrate conflict of competing principles or guidelines. For example, dual use raises the potential conflict of transparency and sharing with the goal of science in advancing the public good. The conflict is translated into competing parties that become involved. For example, the influenza matter that began our discussion is in the courts in the Netherlands where regulations to control possible dual-use findings to protect the public are in direct conflict with academic freedom and openness. No one scientist can make the decision of how to handle situations. Guidelines, recommendations, and laws invariably emerge and no doubt will need to be revised repeatedly as novel situations and improved information emerges.

17.5.4: General Comments

> Let us return for a moment to the common situation in which research is completed and there is no special dual use or related worry. Here the default position for an investigator is to place data and materials used in any investigation in the public domain in keeping with the values of science.

Apart from the fact that sharing reflects a core value, there are scientific benefits of data sharing as well. A given data set can represent a rich resource.

The interests and creativity of a given investigative team that obtained and analyzed the data may not exhaust the knowledge available from that data set. Other investigators with novel hypotheses, varied conceptions of the research issues and underlying theory, and different training or orientations may extract new knowledge. The full potential and yield from a data set may be greatly enhanced if data were routinely shared. Indeed, there is a remarkable untapped resource in data that have been collected but have not been fully mined. The increased sharing of large databases and the pooling of data reflect increased recognition of this potential and to take advantage of data we already have but are not using very efficiently.

Sharing data might be very useful as a general practice to optimize the knowledge that is gained from any study and the combinations of data sets among interested parties. It would be quite useful if investigators could routinely provide in some computerized format of the original article, raw data, and codes for the variables so that the data could be analyzed by others. Inevitably, there would be problems (e.g., software incompatibility, incomplete reporting of the codes, and occasional studies and databases that the profession would just as soon forget rather than re-analyze [see my dissertation]). Large-scale studies that in many ways represent once-in-a-lifetime data sets are prime candidates for data sharing in this fashion. And many such databases are available as I illustrated earlier. Yet, more might be made of individual studies where others might devise novel hypotheses and do not have to collect new information to test them (Hussong et al., 2013).

There may be ethical constraints for sharing and pooling of all data, given that some institutional review committees restrict use of data for the specific purposes outlined in the original proposal and by the investigator who provided that proposal. The purpose is to protect subjects whose consent does not extend to use beyond the original project. Also, investigators often are wary to share data until they have completed their analyses or are protective in general. Making data more available is become more common as journals and agencies require that. The broader ethos is changing too in recognition that data sharing is better for science because of the range of questions that can be addressed. Data sharing is also likely to be a more efficient use of public funds that support much research by maximizing the yield from data that have been collected.

17.6: Conflict of Interest

17.6 Examine how conflict of interest may emerge in scientific research

A critical issue that has emerged in scientific research pertains to conflict of interest.[6] The conflict refers to any situation in which an investigator may have an interest or obligation

that can bias or be perceived to bias a research project. The conflict comes from entering a professional role as a psychologist, but the person has some competing role (personal, legal, financial) that could be expected to impair their objectivity or judgment (APA, 2010a). Among the issues involved in conflict of interest is the potential bias in the findings that are obtained or the position an investigator advocates professionally. Apart from introducing bias, the conflict can undermine the credibility of science. That is, the public believes an objective or objective as possible evaluation was provided only to learn that the scientist has something to gain personally by the direction of the findings.

Impetus for much of contemporary concern reflects research where the investigator may have a financial interest in the results of an investigation.

> The most common example would be if the research is supported by a pharmaceutical company (e.g., grants, consulting fees) and there is an incentive for the investigator to obtain findings or take a position that supports the company.

Research findings can have important financial implications for the investigator in other ways. For example, investigators, especially in the areas of technology (e.g., new software, "apps," robotics) and biology (e.g., new model of drug action, gene therapy), might begin a company and want the rights for commercial exploitation of the findings. Here the investigators are at once scientists and entrepreneurs and the roles clearly can conflict. All sorts of potential conflicts can arise such as whether a faculty member paid by the university should be able to utilize time and resources for an outside commercial interest and whether students working on research are doing so for their educational experience or to contribute to some financial goal (Cech & Leonard, 2001).

In psychological research and perhaps specifically in clinical, counseling, and educational psychology, it is easy to envision conflict of interest. Researchers may own stock in companies that in some way are relevant to their research and their findings. Also, a researcher may serve as a consultant to a company (e.g., that develops software or psychological tests or that publishes textbooks) and receive generous consultation fees for serving as a resource for the company. Serving as someone who gains financially from a company and who conducts research with products that the company may sell could be a conflict of interest or be perceived to have such a conflict. For example, if a company is developing and publishing treatment manuals, a researcher may be solicited to conduct some type of research on these manuals (e.g., surveys of clinicians and what they might like or treatment outcome studies using the manuals). Finding that the therapy works so it could be promoted in textbooks and workshops might lead to genuine conflict of interest or the appearance of conflict of interest, i.e., objective reporting of outcome data by a dispassionate scientist versus reporting of the data by a passionate psychologist who could use or would like the money. Insofar as the researcher stands to gain financially (e.g., stock, further grant funds), the research and consultant roles represent a conflict of interest.

There is a dilemma for researchers:

- Conducting a program of research often requires extensive research funds.
- It is often the case that one grant or funding source will not support an active research lab.
- A major task for senior researchers is to maintain funding for staff and graduate students. Several grants may be sought and any public foundation, resource, or company that funds research is reasonable to pursue.
- Often companies or businesses that have a vested interest in a specific outcome are viable options to obtain funds.
- Without mischief on the part of companies or investigators, it is easily the case that funds are provided by the company with the hope of obtaining results that will help the company. Investigators may have their own agenda such as complying with that, just doing good science, advancing their careers, or putting themselves in a position where they may obtain more funding.

The most familiar context for conflict and perhaps closest to research in clinical psychology relates to the treatment of mental and physical health. In these instances, money is often involved in some way. Investigators might receive grant funding from a company that has interest in showing something is effective, receive consulting fees for their own personal use or fees for their research, or receive stock in the company. The financial issues can be large for a company too; a positive or negative finding on a drug that looked promising actually increases or decreases the value of the stock for that company (Rothenstein et al., 2011).

For example, an investigator may be evaluating the effects of a medication (for obesity, tic disorder, or depression) and have a grant from a company that produces the medication and, of course, would very much like for that medication to be effective. Effective medications for medical and psychiatric conditions, when successful, can earn billions of dollars for a company. This situation could easily represent a conflict of interest for the investigator. The conflict is that science favors careful design of the study and impartial evaluation but the funding agency favors demonstrating an effect. The investigator is on both sides of this, i.e., a conflict, because her funding (grant) and perhaps personal funding (money from the grant or separate consulting fees) might favor a less objective evaluation.

Is there a "real" conflict, and will the findings be influenced by the funding agency?

I have mentioned instances already in which findings funded by companies and industries tend to be disproportionately favorable to the company's or industries product (e.g., soft drinks, tobacco products). Yet, that is not the only issue. Conflict of interest refers to a case where there can be a real conflict or the perception of such a conflict.

The potentially competing interests of companies and the research findings of their product occasionally have dramatic examples where the conflict is in the public media, science media, and the courts. One such illustration emerged in a large-scale study mentioned previously in which a new drug was added to standard treatment to prevent the progression of HIV (Kahn, Cherng, Mayer, Murray, & Lagakos, 2000). All patients (>2,500) from many different sites (77 hospitals) received standard medications designed to lead to lower levels of HIV. Some patients also received a new medication; others received a placebo—all in addition to the standard treatment. The study was stopped early because it was clear that the new drug was not helping at all (e.g., in deaths and progression of HIV). The investigators published the results of the trial (no-difference finding). This led to a major conflict and litigation.

The drug company that sponsored the trial did not want the results published, did not agree with the results, and stood to lose a great deal by publication of the findings (see Burton, 2000). The investigators said they did not agree to have the company control publication, although the company could review the findings before their publication. What the actual contract said between the company and the investigators about publication rights is not readily available public information. The conflict between the company and investigators is not just a minor disagreement. The lawsuit against the investigators sought several million dollars for damages (harm and lost revenue) due to the study's publication.

The conflict here pertains to who has access to the data and the conflict of interest in publishing the results. How this is addressed in research is decided at the beginning of a study when funds are provided. The investigator must work out the details in advance. As an illustration, I have a "friend" who is a clinical psychology faculty member and looks very much like me and in fact is identical in weight, height, rich thick scalp (in place of hair), and a few other features. He worked on a contract/grant as the principal investigator for a government agency. The contract was redundantly explicit in stating that he could not publish or disseminate the results in any form and under any circumstances forever (or maybe even longer) without the explicit written approval of the agency. Data were defined in many ways (e.g., materials developed with the project, reports, charts, drawings, analyses, printouts, notes, and any document finished or unfinished). The contract went on to convey that all information from this project was the property of the agency, was completely confidential, and could not be released. In short, the contract was very clear that my friend has no rights with regard to these data and publication. My friend had to sign the agreement before the funds were provided. Of course, such a request is inherently against the values of science (e.g., transparency, honesty) because all information was not available for public use. My friend has a prison phobia and hence did not violate the contract.

Perhaps the agency was wise in restricting scientific freedom for its own interests. At the end of the project, the evaluation found that services provided by the agency on a large scale were not having impact and one could readily question why the agency continued what they were doing, i.e., spending public money on services that looked like they did not really help people in need. (My friend was vague about the agency, intervention, clientele, but was less vague about the impact of the study on services—"nada" was the word he used, for those of you who understand French.) Unfortunately, the results could not be published or shared in light of the complete control that the agency required in advance. One can understand why—the press and public would have had a feast. Dramatic headlines are easy to envision, "State Spends Millions but No One Is Helped," or "Mental Health Services in the State of . . . Expensive and for What?" More dramatic headlines from specialists in that skill could imply that no one was ever helped and that all treatment cost a fortune. The details would be inaccurate, but the thrust would have led to investigations, more bad press, and so on. What is the outcome of all of this? Years later, the clients continued to receive the interventions that arguably were shown to be ineffective from one evaluation (not replicated). What is the conflict of interest? The state wanted an evaluation but really did not want any news that might be unfavorable.

Another conflict of interest focused on researchers in child psychiatry working with attention-deficit/hyperactivity disorder and then later bipolar disorder in children (see Kaplan, 2011). This case received enormous attention because it involved well-known researchers (see reference for all involved), world-class research institutions (Harvard Medical School, Massachusetts General Hospital), and a U.S. Senate investigation of the case. The researchers had not disclosed their income from pharmaceutical companies. The lead researcher disclosed receipt of $200,000 rather than the alleged $1.6 million he actually received. A 3-year investigation revealed violations related to conflict of interest and not adhering to university or hospital policies. Several penalties were imposed by the university and hospital including requiring the researchers to write a letter of apology to the rest of the faculty,

prohibition of participating in activities that produced income from pharmaceutical companies for one year, and requiring that formal permission be sought after that year for participation in such activities. Failing to disclose information is not a minor issue.

17.6.1: Procedures to Address Conflict of Interest

Many procedures are in place to address conflict of interest. Professional organizations generally advocate for or require very clear statements by the investigator if there is any possible conflict of interest. Occasionally the recommendations include stating in the informed consent form any potential association of the investigator that is or could be conceived as a conflict. Also, in any research publication, funding sources or possible conflict of interest is to be mentioned, sometimes in the letter that accompanies the manuscript when it is submitted for publication and then again in a footnote in the article itself. Universities that receive federal research funds are mandated to ask faculty to disclose any conflict of interest they may have. Faculty are asked whether they own stock in a company or have a significant financial income from that company (e.g., earns or receives more than $10,000 per year or has stock or related ownership interest more than $5,000) or if their research is supported by a company or organization that might provide or appear to provide a conflict of interest.

> Formal agreements usually need to be signed by a faculty member annually, and faculty are required to update their conflict of interest statement (usually a form filled out online) if the status of the conflict of interest changes (e.g., by being a consultant or on the board and now receiving money that triggers some arbitrary number of constituting a conflict).

If an investigator does have a conflict of interest as defined by the regulations and policy, some actions are taken to mitigate this in some way such as asking the investigator not to be involved in decision making related to the funding agency or source, abstaining from activities that cause the conflict, or close monitoring of the activities to help in some way to reduce bias. The main and most common intervention about managing conflict of interest and its appearance is requiring public disclosure on the part of the investigator.

Not all conflict of interest is financial, although that is the main concern in federal law and policy regarding federally funded research in the United States (see http://grants.nih.gov/grants/policy/coi/). Sometimes intellectual conflict of interest is distinguished from financial conflict of interest, although they are related. Intellectual conflict of interest refers to academic activities in which the investigator is connected with a particular point of view that could affect his or her judgment (Guyatt et al., 2010). That conflict may arise from some financial connection but goes beyond any particular study or report. Rather, the individual's broader position or judgment might be seen as unduly influenced by a connection with a particular company, business, or other entity. That influence and commitment now can be seen as shading judgment, evaluation of the data, and recommendations that may stem from that. This is an iffy area and topic because one cannot tell if the judgment, evaluation, and recommendations are based on the researcher having a conflict of interest or having an unbiased view that coincidentally is keeping with what a conflict of interest might suggest. Scientists often differ in their views and recommendations and interpretation of the evidence. Thus, one cannot tell if a particular view represents a true conflict of interest. That is why the appearance of a conflict or potential conflict is in professional guidelines as "counting" as a problem. The time line may even vary from what we assume. Namely, the researcher's intellectual view about some phenomenon may have occurred without any conflict of interest and that view led her to be sought by companies.

17.6.2: Other Conflicts of Interest Briefly Noted

Conflict of financial interests and the research enterprise go beyond the individual investigator. Consider a few briefly:

1. Major research universities often have resources devoted to assisting investigators in launching start-up companies that are for profit. Research often reveals a treatment (e.g., medication for a psychiatric disorder), procedure (e.g., to study a biological process), or technological advance (e.g., new type of solar panel) that may lead to a patent and to a commercial product. The goal is to utilize findings obtained in research to benefit the public. This transfer of technology often has been completed by business and industry (commercial companies), and universities have entered into this to benefit from the gains. Actually, the initial impetus is to move a product from the lab to the community (of researchers or the public), and businesses do this better than universities. Universities may have offices that facilitate this process and even provide start-up costs and direct assistance. Universities often take a percentage of the funds when a product has been developed under their roofs so to speak. Thus, universities have some potential conflicts too because their earnings can come from a particular line of research or set of researchers. Vested interest in financial gain of universities is not usually part of the discussion of conflict of interest. Perhaps this might not be regarded as a conflict in the sense that both the public good and financial gain may operate in the same direction, i.e., call for

moving research to application. Also, universities do not oversee a particular project or root for the results one way or another on that project.

2. Journals often have a conflict of interest in which their goal of publishing the best science competes with another goal of making a profit. There are now hundreds of online, open-access journals that charge authors for publication. Authors submit their manuscripts, the manuscripts are reviewed (or not), and accepted for publication; then charges are billed to the investigator.

   The conflict—some of the journals are not very interested in science or the science standards. They are for-profit journals that make money on the basis of how many articles they accept. Author fees are the source of income.

The journals usually can be identified. They have obscure locations of Web sites, often have no academic affiliation, avoid disclosing their manuscript review procedures, initially hide the fact that they will charge fees to the author to publish the manuscript, and may have fictitious people on an "editorial board." They are sometimes called *predatory journals* (see Beall, 2012, for a long list of criteria for being so classified). Among the criteria, it is difficult to locate or find the publisher or editor. Searches (for computer IP address) often reveal locales of the author in obscure countries or locations even though the editorial information suggests a Western country where research practices and peer review are well developed. Thousands of open-access journals are of this type and span the full range of scientific disciplines. Again, the conflict is at the level of the journal. Publishing solid science is not the goal; money making is. The entire operations are based on deception. Moreover in many cases, the usual standards of scientific integrity (no publication of the same paper in two outlets) are not required. The journals are not all that hard to identify by someone actively involved in research in an area related to the journal's title. Yet, they also raise other issues such as another source of mistrust and incredibility of science that filters to the public. Virtually anything can be published and that could be a finding that enters its way into the media. Am I exaggerating? One investigator developed a spoof treatment study that was purposely designed to be of horrible quality and one that "reviewers would easily identify as flawed and unpublishable" (Bohannon, 2013, p. 62). The article with slight variations was submitted to 304 open-access journals as part of a systematic study. Over 50% of the journals accepted the manuscript, often with accolades and invariably with publication fees if the author decided to go through with actual publication.

Related, there are also *predatory conferences*. These are usually fake international conferences where a researcher receives an invitation to present a paper at some world congress or international venue. The conference looks legitimate and may even have some fake invited world leaders already presenting. Here too as with the journals, there usually is no "real" professional organization, agency, or even single scientific discipline associated with the conference. Also, the conference e-mails convey that one is invited to deliver a paper or even a special address. The invitations are often bizarre. (In the last month as I write this, I have received three invitations: one to talk about plants and agriculture, another on engineering, and another on nanotechnology. Even at my most grandiose moments when I believe my research could solve 90% of the world's problems, I usually leave out these three areas!)

University financial interest in products of research, predatory journal publishing, and predatory conferences are fascinating topics in their own right. Yet, I mention them in passing because some do actually involve integrity issues for the individual researchers (e.g., publishing and presenting in predatory places—be careful). However, the main reason was to convey that conflict of interest in scientific research usually refers to individual investigators, an emphasis I provided earlier. It is merely useful to note that conflict of interest and good science versus financial gain are not just with the investigator.

17.7: Breaches of Scientific Integrity

17.7 Identify instances that cause breaches in scientific integrity

Science and those who oversee various facets of research (e.g., granting agencies, professional organizations, faculty mentors of younger colleagues or students, journal editors) are deeply concerned about lapses in ethical issues and scientific integrity.

17.7.1: Jeopardizing the Public Trust

Overall a uniting theme of the sciences is to improve the world and public life (quality, safety, health) and to sustain conditions (e.g., climate, ecosystems, habitats) that support that. That is undermined when there are breaches of the public trust. A healthy skepticism of any scientific finding, in my view, is to be actively encouraged. We do not change our diet and bedtime ritual because one or even two studies find that ground oatmeal and chia seeds, taken intravenously seconds before going to bed, lead to a longer life, reduced depression, and boundless energy. These clichés and questions are all important and central to what we do as scientists, i.e., they are legitimate:

- "Too good to be true."
- "Show me more data."

- "Has the finding been replicated in well-controlled studies?"
- "Is there any company or investigator behind this that might profit from us believing the finding?"

In short, *skepticism* is fine and indeed often appropriate. Lapses of scientific integrity are more likely to lead to *cynicism*, which is quite different. Here the public distrusts and sees science as just another place where self-serving individuals are promoting a position and not providing "facts" in any way.

Trust is difficult to earn and easy to lose. Witness the vaccination-autism episode I mentioned. Long after the scientific record has been corrected, long after careful studies have shown the original claims were wrong, and long after prestigious panels of leading experts claimed vaccinations do not cause autism, we are still in a distrust phase where many individuals and many efforts (e.g., Web sites) still foster suspiciousness of science and scientists. And many children are not being vaccinated. It is easy to comment that "my research is not that relevant or important and could not hurt anyone in that way." But of course that is not the issue. Breaches of scientific integrity bring down the whole enterprise, and all are stained by it. Also, of course harming the public trust in an area of research that does not affect them directly could foster distrust of some other finding or practice that does affect them.

Prior to starting a research project, the investigator must have a proposal approved by an Institutional Review Board. The salient focus is on the issues related to the protection of the participants, as evident by considering if deception is used, whether the consent procedures are appropriate and comply with federal regulations, how privacy information will be protected, and other such issues. There are no analogous protections of scientific integrity that evaluate a project and check to ensure that there will be no lapses of integrity. Lapses of integrity often are after a study has been completed (e.g., plagiarism, inappropriate or misallocation of credit) or completed behind closed doors (e.g., fraud).

Once a study has been completed if there are suspected violations of scientific integrity, universities have procedures to investigate them and invoke consequences. Often matters are also turned over to the criminal justice system as relevant (e.g., misuse of federal funds, violation of HIPAA). Universities vary in how quickly, how decisively, and strongly they respond to allegations of fraud or scientific misconduct. That delay can readily be interpreted as reflecting little or no interest in responding to violations of scientific integrity. That interpretation may be true or partially true because once revealed everyone is stained in the process, the investigator, the university administration, and the name and value of the university itself. The consequences can even translate to money as donors are reluctant to contribute to a shamed university. Also, scrutiny of one instance of something might reveal more instances of other things that are also suspicious. There is an investigator or reporter somewhere who is eager to write a story that notes the violation of this scandal is "not the first time the university has done x or y." Yes, there might be real incentives for a university to drag its investigative feet.

On the other side, there are complexities of investigating fraud, collecting the information, interviewing witnesses from the lab from where the suspected practices emerged, preparing a report, and so on. Throughout the process, those involved in the investigation are thinking litigation in two ways:

1. Making missteps that might cause the university to be sued
2. Keeping options open for suing others

The university does do not want to trample anyone's rights, make moves that could be misinterpreted, or jeopardize positions (e.g., university presidents, boards), and research funds at the university.

Often too there are innocent victims (e.g., postdoctoral researchers, graduate students) who were not directly involved but will suffer (loss of positions in the lab, delay of graduate theses). Thus, caution and a measured pace usually characterize evaluations of scientific integrity.

17.8: Remedies and Protections

17.8 Determine remedies and protections to safeguard ethical interests of the subjects of statistical research

After the fact investigation of breaches of scientific integrity is necessary part of the process of evaluating, judging, and if necessary punishing lapses of integrity. Yet, after the fact investigation is not the main emphasis or procedures. Through the chapter, I have many proposed solutions to individual problems. It is worth highlighting many strategies to convey that while there are occasional lapses that are significant there are also many remedies and protections in place to ensure ethical treatment of participants and high levels of scientific integrity.

A widely accepted view is that education of researchers is the best strategy to prevent lapses in ethical care and scientific integrity. Many resources are available that convey the guidelines from several organizations, and these can serve as a resource (see http://grants.nih.gov/grants/research_integrity/).

A prominent example is the Office of Research Integrity (United States Department of Health and Human Services, 2012), which is a resource for policies, accumulation of findings, and cases about misconduct. It provides

assistance to universities in handling allegations of misconduct. Importantly, it provides guidelines and training materials that can be used with students and faculty to promote research integrity and prevent misconduct.

Guidelines alone do not ensure adherence to research responsibilities. Consequently, a key issue is how to ensure that persons involved in research are exposed to guidelines and the key topics. Accreditation of training programs (e.g., in clinical and counseling psychology) requires exposure of students to ethical issues and guidelines. More generally, universities involved in federally funded research must ensure that all persons involved in the research (principal investigators, postdoctoral researchers, students, and assistants at all levels) are exposed to some university-based instructional program (classroom, Web based) that discusses responsibilities of the researchers, informed consent, conflict of interest, publication practices, authorship, data sharing, and related issues. Instruction has now become a matter of policy for institutions involved in federally funded research (see http://ori.dhhs.gov/html/programs/finalpolicy.asp). Also, investigators have to explicitly attest to completion of training, often on an annual basis, and convey any potential conflict of interest. In short, there are educational materials and training opportunities provided to researchers and mandatory activities to be sure that researchers know the rules, regulations, and accepted practices.

All sorts of changes have been made in publishing of scientific articles to help address scientific integrity. For many journals, author contributions have to be made explicit in multiauthored papers, and the data underlying the study may need to be deposited or made available for use by others.

> Many journals are trying to give greater attention to so-called "negative results" (no statistically significant differences) because the publication bias against such findings in part fosters and unwittingly provides strong incentives for researchers to look for and find significance, as discussed in relation to both fraud and questionable research practices.

Emphasis on replication of findings has gained considerable momentum with special initiatives in biomedical sciences and psychology (e.g., Carpenter, 2012; Couzin-Frankel, 2012). The airing of these initiatives in various science journals, newsletters, and blogs no doubt will spread interest in fostering replications and their publication. We want to be sure that our findings in fact can be replicable and are stable. Also, we want procedures, methods, and data to be shared so that studies can be replicated. The relative ease of storing extensive material electronically makes storing of materials possible. The availability of materials and access to them by more researchers will allow checking on findings.

I have highlighted several protections to convey that much has been and is being done to ensure that standards of ethical behavior and scientific integrity are achieved. There is much at stake that can undermine public trust, can lead individuals (public and other scientists astray with false information), and more. Leading the public and other researchers down one path necessarily utilizes resources (funds, scientific talent) that could have been deployed elsewhere. In short, scientists and policy makers are working on all of this and indeed many professionals (e.g., within psychology, other sciences, law) have career paths that focus primarily and often exclusively on the ethical and scientific integrity issues.

With all remedies and protections in place, it is important not to lose sight of givens of science. To begin with, scientists are human. Thus, the full panoply of human characteristics, motives, and foibles is likely to be evident. All the characteristics that make a Shakespeare tragedy, mystery novel, and television sit-com intriguing and interesting can spill over in some way into science. This does not mean that the negative virtues are pervasive or that one ought merely to shrug one shoulders and say "of course, what did you expect" whenever a lapse in ethics or scientific integrity occurs. It does mean that we should not be shocked to hear instances when less-than-desirable samples of humanness are evident, as when researchers argue ad hominem about their theoretical differences, when beliefs are held to tenaciously in the face of seemingly persuasive of counter evidence, or when differential standards are applied to interpretation of some kinds of work (mine) rather than other kinds of work (yours). As humans, we are by our very nature limited and the area of cognitive heuristics and decision making are merely two broad areas of psychological research that illustrate "normal" biases in operation. We are motivated viewers; we bring subjectivity to our experience and its interpretation. Fraud, interests in primary credit, possessiveness of ideas, procedures, and data, as discussed previously, occur and hence always warrant attention.

That scientists are human does not excuse lapses that compete with the very purposes of science. In discussing the lapses, we ought not to lose sight of the other, positive side. Humans have invented science and all of the methods, procedures, practices, and values aimed at increasing objectivity and replicability. There is an enormous commitment, curiosity, and integrity among professionals in all of the sciences to discover, understand, and reveal. Subjectivity, error, and bias cannot be eliminated. Indeed, some of the very methods used to decrease subjectivity introduce their own sources of error, artifact, and bias. For example, statistics are used to provide a criterion to determine if there is a reliable effect; yet chance, in any given case, could explain the difference. As likely in much of research, the absence of differences can be an artifact of weak power. Also, measures are used to permit evaluation of constructs and to provide more objective means of assessment than impressions and personal opinions of the investigator;

reactivity of assessment and low validity of the measure are potential limits assessment often introduces. However, these sources of error can be placed within the scientific arena, investigated, and evaluated.

Science plays a critical role in society and if anything that role has expanded in recent years. Entirely new topics emerge, new hybrid sciences take shape, and novel methods of assessment reveal new levels of analysis. For example, the entire brain can be scanned rather than sections looking for activation here and there. We know now that the microbes in the human body vastly outnumber the cells that form our bodies and that these microbes somehow are involved in learning, memory, immune response, and more. That more no doubt will involve psychological states and functioning.

As science continues so will procedures and practices continue to monitor and ensure adherence to ethical issues and scientific integrity. Scientists more than any other group realize the importance of maintaining the integrity of what we do. We welcome scrutiny because transparency, openness, and accountability are part of the core values.

Summary and Conclusions: Scientific Integrity

This chapter focused on scientific integrity and the obligations and responsibilities of investigators to maintain the core values of science and carry out the practices with which these are associated. *Ethical issues and scientific integrity* form critical components of the research and emerge at all stages and steps of research from developing the proposal and obtaining approval to proceed through the data analysis, write-up, and publication of the final product.

Many critical issues of scientific integrity were discussed. We began with core values and included transparency, honesty, accountability, commitment to empirical findings, addressing or avoiding conflict of interest, and commitment to the public's interest. These are a useful starting point to convey what underlies what we are doing and how we go about the business of research.

Core values help guide many specific topics that are specified further in ethical guidelines (e.g., American Psychological Association), policies, regulations, and federal law in the United States. Several specific topics were discussed in detail, including fraud in science, questionable practices in research, plagiarism, allocation of credit to collaborators, and conflict of interest. Many concepts were introduced along the way, including honorary or gift authorship, ghost authorship, and self-plagiarism.

One concept that was introduced was "big data," which is a new emphasis in many of the sciences. Big data essentially refers to massive amounts of information that are now available and how that can be integrated and used to make novel advances. The concept is useful to convey some of the new challenges in ethical issues and privacy that are raised by advances in science. There are many guidelines and regulations to address ethical issues and scientific integrity. Big data is a useful illustration of how guidelines and regulations always need to be evaluated to keep up with new situations, concerns, and potential ways in which participants and scientific practices may need to be protected.

Science is designed to serve the public. Our understanding of phenomena is to increase the knowledge base in ways that will improve the conditions of the world and living inhabitants. That is a huge challenge and responsibility and makes ethical issues and scientific integrity critically important. There are many protections in place to minimize lapses in ethical behavior and scientific integrity and many remedies once such lapses are identified. And these are constantly being revised to keep up with any new circumstances. Education of budding scientists and continued education of those well into their careers are some of the basic elements in transmitting the values, practices, and responsibilities of scientists.

Critical Thinking Questions

1. Most research not only in psychology but also in natural, biological, and other social science has no immediate application. Even so, fraudulent reporting of findings could still jeopardize public trust and cause the public to stop engaging in a scientifically based practice that does affect personal welfare and health. How could this happen? Give a real or hypothetical example.
2. Plagiarism is a problem on university campuses among undergraduate and graduate students. Identify two or three effective ways that might help combat that.
3. Conflict of interest of investigators is handled by asking them to disclose their sources of conflict (e.g., before presenting their work at a talk or in an article). How might this be helpful or effective in addressing the conflict? How might it not be helpful?

Chapter 17 Quiz: Scientific Integrity

Chapter 18
Communication of Research Findings

Learning Objectives

18.1 Recognize the importance of informative and clear scientific writing

18.2 Show the outline of creating a robust manuscript

18.3 Report the general sections that should be a part of scientific writing

18.4 State the primary goal of robust scientific writing

18.5 Identify guidelines of creating a successful scientific writing

18.6 Recognize the importance of selecting the appropriate journal for scientific publication

18.7 Detail the scientific publication submission and review processes

The research process is composed of the design, execution, analysis of the results, and preparation of a report. This "report" is the way to communicate findings. In professional academic life, the report usually is for a journal article, for a presentation at a convention, or for an abbreviated poster session where the study is summarized on one large poster type sheet for others to review as they walk through a convention hall. For students in training, the study may be written for a course or thesis project (senior or master's thesis, doctoral dissertation). And communication of one's findings can be directed to different audiences (e.g., other professionals within one's field, the science community more broadly, and the public and media). Each format and audience has its own nuances.

In this chapter, I emphasize communication of findings through the write-up of the results of a study for journal publication. Focus on journal publication is useful in the chapter because this allows for discussion of the interface and connections of methodology with communication of one's findings. Also, a critical goal in science is to disseminate one's work in an archival source and journal publication (e.g., more than textbooks and chapters) is the usual format. Key issues in preparing a report for publication including how to present the study, the rationale, and other information apply broadly to reporting on the results of a study even when publication is not the goal. Thus, theses and dissertations, for example, like journal articles raise similar challenges, namely, how to present the research in its best, clearest, and also most persuasive light. As it turns out, knowledge of methodology as well as expertise on the topic can help in preparing reports of one's study.

This final step of writing up an article for journal publication seems straightforward and relatively easy, given the nature and scope of the other steps and after all we have been through just to get the study done. In fact, one often refers to preparation of the article as merely "writing up the results." Yet the implied simplicity of the task belies the significance of the product in the research process and the challenges in making the case for why the given report of a study ought to be published. In addition, there is a sense in which the manuscript is not the final step in the research process. Rather, it is an important beginning.

> The article is often a launching platform for the next study for the authors themselves and for others in the field who are interested in pursuing the findings.

Thus, the report is central to the research process.

Publication of research is an essential part of science and the accumulation of knowledge. That accumulation requires ways to archive the studies, so present and future researchers and others (e.g., policy leaders) can draw on them. Related to the accumulation of knowledge is dissemination of one's findings. That is, we do not only want the study archived in some dusty shelf or some long-lost pdf files in the bowels of

the Internet. We want the results to be circulated, perhaps to address a critical question or to influence other researchers. Publication can serve other goals as well. Many professional and career goals served by publishing one's research.

Publication of one's research can signal a level of competence and mastery that includes:

- Developing an idea
- Designing, executing, and completing the study
- Analyzing the results
- Preparing a written report
- Submitting it for publication
- Traversing the peer-review process

This chapter focuses on publishing one's research as a primary way of communicating results. The thinking and organizing the information for publication have broad generality in preparing reports for other purposes. Publication has its own special processes and challenges, and we will take these up as we discuss preparing a manuscript, selecting a publication outlet, submitting the manuscript for review, and revising the manuscript as needed for publication.

18.1: Methodologically Informed Manuscript Preparation

18.1 Recognize the importance of informative and clear scientific writing

A central goal of scientific writing is to convey what was actually done so that the methods and procedures can be replicated. *Concrete, specific, operational, objective, and precise are some of the characteristics that describe the writing style.* The effort to describe research in concrete and specific ways is critically important. However, the task of the author goes well beyond description.

18.2: Overview

18.2 Show the outline of creating a robust manuscript

Preparation of the report for publication involves three interrelated tasks that I refer to as:

- Description
- Explanation
- Contextualization

Failure to appreciate or to accomplish these tasks serves as a main source of frustration for authors, as their papers traverse the process of manuscript review toward journal publication.

Description is the most straightforward task and includes providing details of the study.

Even though this is an obvious requirement of the report, basic details often are omitted in published articles (e.g., sex, socioeconomic status, and ethnicity of the participants; means and standard deviations).[1] Omission of basic details can hamper scientific progress. If a later study fails to replicate the findings, it could be because the sample is very different along some dimension or characteristic. Yet, we cannot surmise that without knowing at least basic details of the sample in both studies. If a study does repeat the findings, that is important but is the new finding an extension to a new type of sample? Again, we need basic information in the studies to allow such comparisons.

Explanation is more demanding in so far as this refers to presenting the rationale of several facets of the study.

The justification, decision-making process, and the connections between the decisions and the goals of the study move well beyond description. Here the reader of the manuscript has access to the author's decision points.

There are numerous decision points pertaining to such matters as:

- Selecting the sample
- Choosing among many options of how to test the idea
- Selecting the measures
- Including various control and comparison groups

The author is obliged to explain why the specific options elected are well suited to the hypotheses or the goals of the study. There is a persuasion feature that operates here. The author of the manuscript is persuaded that the decisions are reasonable ways to address the overriding research question. Now the author must convey that to persuade the reader. In other words, explanation conveys why the procedures, measures, and so on were selected, but that explanation ought to be cogent and persuasive. We do not want the reader to think, "This is an important research question, but why study it that way?" For the many decision points, that very reasonable question has to be anticipated and pre-empted.

Finally, *contextualization moves one step further away from description and addresses how the study fits in the context of other studies and in the knowledge base more generally.*

This latter facet of the article preparation reflects such lofty notions as scholarship and perspective, because the author places the descriptive and explanatory material into a broader context.

Essentially, the author is making the case for the study based on the knowledge base.

Relatively vacuous claims (e.g., this is the first study of this or the first study to include this or that control condition or measure) are rarely a strong basis for the study and often

means or are interpreted as meaning that the author could not come up with something better. Without context, any "first" is not very important by itself. Indeed, it is easy to be first for a topic that is not very important and has been purposely neglected or relegated to a very low priority. We need a more compelling rationale.

For example, if this study is done on why people commit suicide, we need the context of why this particular study ought to be done and where in the puzzle of understanding this piece fits. Perhaps prior research omitted some critical control procedure, perhaps there is a special group that has a novel characteristic that reduces (or increases) the likelihood of suicide that would inform the field in unique ways, or perhaps some new twist on a theory or intervention will have clear implications for reducing suicide attempts. These and other such comments convey there is a gap in knowledge, that gap is important, and that gap will be filled in whole or in part by this particular study. Among researchers beginning their careers or early in the publication game, contextualization is likely to be the greatest challenge.

The extent to which description, explanation, and contextualization are accomplished increases the likelihood that the report will be viewed as a publishable article and facilitates integration of the report into the knowledge base. Guidelines are provided later in the chapter to convey these tasks more concretely in the preparation and evaluation of research reports. The guidelines focus on:

- The logic of the study
- The interrelations of the different sections of the manuscript that describes the study
- The rationale for specific procedures and analyses, the strengths and limitations, and where the study fits in the knowledge base

Consider main sections of the manuscript that are prepared for journal publication and how these components can be addressed.[2]

18.3: Main Sections of the Article

18.3 Report the general sections that should be a part of scientific writing

Here are the components of an article to be submitted for journal publication.

18.3.1: Title of the Article

Well, certainly the title is *not* a "main section" of the article, but it is not trivial either. The title may determine whether a potential reader of the article goes on to the Abstract and rest of the article.

> Usually one attempts to address the key variables, focus, and population with an economy of words.

If the study focuses on diagnosis, assessment, treatment, or prevention, one of these words or variations might well be included. Similarly, if a specific disorder (e.g., depression), personality characteristic (e.g., repression-sensitization), treatment technique (e.g., structural family therapy), or sample is critical (e.g., infants, elderly), the pertinent terms are likely to be integrated into the title. Similarly, any salient characteristic of the focus (e.g., emotion regulation, subjective distress, biomarkers, neuroimaging, attention deficits, and rumination) ought to be salient in the title.

It is critical here to be direct, clear, and concise (e.g., "Memory loss and gains associated with aging" or "Predictors of drug use and abuse among adolescents" or "Trauma symptoms among veterans who have not seen combat"). These examples are especially concise as well as clear. On the other side, try to avoid vague or ambiguous terms. Examples might be terms such as an "at risk sample" (at risk for what?) or "parenting influences on child school behavior" (what parenting influences and what school behavior), and "memory deficits as a function of interpersonal violence" (what kind of memory—many different types in psychology and what kind of interpersonal violence?). We are only in the title section of the manuscript. It would be nice not to reveal this early in the write-up of the manuscript that our thinking is fuzzy, we have no clear focus, and key goals or concepts are vague.

Ordinarily an author is encouraged to fit the title within 10–12 words. The words ought to be selected carefully. Titles occasionally are used to index articles in large databases. Words that are not needed or that say little (e.g., "preliminary findings," "implications," "new findings") might be more judiciously replaced by substantive or content words (e.g., among preschool children, the elderly; consequences for sleep and stress) that permit the article to be indexed more broadly than it otherwise would have been.

Occasionally, comments about the method are included in the title or more commonly in the subtitle.

> Terms like "a pilot study" or "preliminary report" may have many different meanings, such as the fact that this is an initial or interim report of a larger research program.

These words could also be gently preparing readers for some methodological surprises and even tell us not to expect too much from the design. (For example, my dissertation coined the subtitle: "A pre-preliminary, tentative, exploratory pilot study©.") In some cases, terms are added to the study, such as "A Controlled Investigation," which moves our expectation in the other direction, namely, that

the present study is somehow well conducted and controlled, and perhaps by implication stands in contrast to other studies in the field (or in the author's repertoire). Usually words noting that the investigation is controlled are not needed unless this is truly a novel feature of research on the topic. Select words carefully and try to make as many words in the title reflect content of what is in the study.

Occasionally authors want to use titles with simple questions, "Is depression really a detriment to health?" or "Is childhood bullying among boys a predictor of domestic violence in adulthood?" In general, it is advisable to avoid "yes, no" questions in the title. Scientific findings often are nuanced, and findings are likely to be both yes and no but under very different circumstances or for some subgroups of people but not for others.

As an example, consider a hypothetical yes-no question for the title of a study as, "Is cigarette smoking bad for one's health?"

For anyone on the planet, the answer might be a resounding yes. Yet, the yes-no nature of the question makes this a poor choice of title because the answer is likely to depend on either how smoking is defined (e.g., how much smoking—a cigarette a year, a pack after each meal) and how health is defined (e.g., mental, physical, what diseases, disorders). Very familiar is how horrible smoking is for one's physical health in so many domains (e.g., heart disease, cancer, chronic respiratory disease), but the question in the title can be answered both yes and no. Less familiar is the fact that cigarette smoking reduces the risk for Parkinson's disease and there are reasonable explanations for that based on brain chemistry and neurotransmitters (Miller & Das, 2007). So the hypothetical title is not very helpful or informative because we can show many circumstances in which yes and no are correct answers to the same question. I am not arguing in favor of cigarette smoking. I am advising against titles of empirical articles that have a yes-no question.

Few phenomena allow the simplistic thinking the question can reflect, and again it is helpful not to reveal so quickly—we are still only on the title—that our own thinking comes down to true-false questions in a world where most things are essay questions. There might well be exceptions, but ponder the title carefully for your own studies. If you are reading the works of others and see a true-false question in the title, try to consider if there might be exceptions to either yes or no.

18.3.2: Abstract

Why so much time on the title? Because that is likely to be the most widely read part of an article with a sharp drop-off on the proportion of people who continue to the Abstract, the next part with its own demands.

> The Abstract is likely to be read by many more people than is the full article. The Abstract will be entered into various databases and be accessible through Internet and online library searches.

Many journals list the tables of contents for their issues and provide free access on the Web to Abstracts of the articles but charge for the full article. Consequently, the Abstract is the only information that most readers will have about the study.

For reviewers of the manuscript and readers of the journal article, the Abstract conveys what the author studied and found. Ambiguity, illogic, and fuzziness here are ominous. Thus, the Title and Abstract are sometimes the only impression or first impression one may have about the study. You may have the most dazzling study that will cause a news media frenzy, endless e-mail requests for TV appearances, Award committees jamming your smartphone trying to reach you to be the first to recognize your brilliance, and paparazzi waiting all night outside your recreational vehicle just for a photo. Not likely to happen if no one reads the study and readily grasps the key findings.

Obviously, the purpose of the Abstract is to provide a relatively brief but full statement of goals, methods, findings, and conclusions of the study. Critical methodological descriptors pertain to:

- The participants and their characteristics
- Experimental and control groups or conditions
- Design
- Major findings

Often space is quite limited; indeed a word limit (e.g., 150–250 words maximum) may be placed on the Abstract. It is useful to deploy the words to make substantive statements about the characteristics of the study and the findings, rather than to provide general and minimally informative comments. For example, vacuous statements ("Implications of the results were discussed" or "Future directions for research were suggested") ought to be replaced with more specific comments of what one or two implications and research directions are (e.g., "The findings suggest that the family and peers might be mobilized to prevent drug abuse among adolescents," "Cultural influences appear to play a major role in onset but not the course of depression"). Also, the more specific comments can convey the study's relevance and interest value beyond what is suggested by the manuscript title or opening comments of the Abstract. I personally am not going to read very eagerly an article with the vacuous "implications" or "future directions" sentences, but if I am interested in the specific topics mentioned as implications (family, peers, culture), this article is a must for me to read. As authors, we often lament the word restrictions placed on us in the Abstract, but the first task is to make sure that we are using the existing allotment with maximum information.

18.3.3: Introduction

The Introduction is designed to convey the overall rationale and objectives. The task of the author is to convey in a crisp and concise fashion why this particular study is needed and the current questions or deficiencies the study is designed to address. The section should not review the literature in a study-by-study fashion, but rather convey issues and evaluative comments that set the stage for the study. A deadly style that will not place the study in the best light is to begin paragraph after paragraph with the names of the authors with one paragraph beginning with, "Lipshitz and Johnson (2011) studied this and found that and then jumping to the next paragraph, Scooby and Skippy (2012) found that also, but only one two of the measures." (I am already dozing.) Most of the time, the names of the investigators are not important information to lead with: Make the intellectual, academic, or scholarly point of why any particular study is a building block in the logic of what you are doing and of course place the names of the authors of the studies at the end of the sentences. Ideally you can make sentences that combine multiple studies. You are making points, arguments, not presenting studies.

After the initial material, the Introduction moves to the issues that underlie this particular study. Here the context that frames the specific hypotheses of the study is provided and reflects theory and research that are the impetus for the investigation. There is an introduction syllogism, as it were, a logic that will lead the reader from previous theory and research to the present study with a direct path. Extended paragraphs that are background without close connections to the hypotheses of the study serve as a common weakness of manuscripts rejected for publication. Somehow the author feels he reviewed the relevant literature and now opens the curtain for the three hypotheses. Yet, the "relevant" literature is not studies on the broad topic but the studies that serve as the bases for the hypotheses. The hypotheses should not be a surprise but rather easily seen consequences of the literature that has been reviewed.

Placing the study in the context of what is and is not known (contextualization) and the essential next step in research in the field requires mastery of the pertinent literatures, apart from reasonable communication skills.

> Ironically, mastery of the literature is needed so that the author knows precisely what to omit from the Introduction.

A vast amount of material one has mastered and that is very interesting will need to be omitted because it does not set the stage or convey the precise context for this particular study.

Saying that the study is important (without systematically establishing the context) and noting that no one else has studied this phenomenon (measure or sample) usually are feeble attempts to short-circuit the contextualization of the study. Among the tasks of the Introduction is to lead the reader to the conclusion that the study is important and worthwhile. Telling the reader that the study is important and worthwhile is more like an argument from authority and that is not how science works at all. Also, that kind of presentation might even suggest that author has not done his or her contextualization homework and cannot really make the case for the study.

> One way to establish the importance of the article is to convey a "tension." That tension reflects competing views, theories, or findings.

To create a tension, four components are helpful to include:

1. Give one side that perhaps theory or available findings seem to support one view (e.g., individuals who engage in self-injury have this or that characteristic). Make that case as strongly as the literature allows and add your own speculations if they are to be tested.
2. Give the other side that seems to be less clear or even better seemingly contradictory. That is, what seems to be different from the one side that was suggested. Find, convey, show a conflict, discrepancy, or different implications from the two views you have provided. Perhaps the first side is generally but not always true. That is one kind of tension because it raises the "why not everyone or most people?"
3. Convey why we should care about this seeming conflict, and why it is important in relation to theory or application. This is critical. Central to an Introduction is to convey why the reader should care about the focus, hypotheses, and finding.
4. Convey what is a possible resolution to the tension—there is some critical third variable perhaps. The possible resolution is the focus of your study.

Now with these four components, we have a problem or tension that remains to be resolved (first and second component) and a statement that this problem is not trivial but makes a difference in our thinking, theory, application (third component). Finally, this study is contextualized so nicely because you convey that the study you are doing in exactly aimed at the problem and is important (first, second, and third components all at once). The components serve as a useful template to consider, but of course may not always apply. Yet, when applicable the template addresses the importance and logic of the study and helps the reader to see the likely contribution.

18.3.4: More Information on the Introduction

It may be relevant to consider limitations of previous work and how those limitations can be overcome. These statements build the critical transition from an existing literature to the present study and the rationale for design improvements or additions in relation to those studies. It is important to emphasize that "fixing limitations" of prior work is not necessarily a strong basis for publishing a study. The author must convey that the limitations of prior work are central to a key building block in theory or the knowledge base. Convey that because of that limitation, we really do not know what we thought we did or that there is a new ambiguity that is important but hidden in prior studies in light of what was studied and by what means. Alternatively, the study may build along new dimensions to extend the theory and constructs to a broader range of domains of performance, samples, and settings. The rationale for the specific study must be very clearly established. Theory and previous research usually are the proper springboard to convey the importance of the current study. But in all cases, do not assume that the importance of your particular study will be easily grasped.

In general, the Introduction will move from the very general to the specific. The very general refers to:

- Opening of the Introduction that conveys the area
- General topic
- Significance of a problem

For example, in studies of diagnosis, assessment, treatment, or prevention of clinical dysfunction, the Introduction invariably includes a paragraph to orient the reader about the seriousness, prevalence or incidence, and economic and social costs of the disorder. Reviewers of the manuscript are likely to be specialists in the area of the study and hence know the context very well. Yet, many potential readers would profit from a statement that conveys the significance, interest, and value of the main focus of the study.

The Introduction does not usually permit us to convey all of the information we wish to present. In fact, the limit is usually 4–5 manuscript pages. A reasonable use of this space is in brief paragraphs or implicit sections that describe the nature of the problem, the current status of the literature, the extension to theory and research this study is designed to provide, and how the methods to be used are warranted. The penultimate or final paragraph of the Introduction usually includes a statement of the purpose of the study and the specific hypotheses and predictions. By the time the reader reaches this paragraph or set of paragraphs, it should be very clear that these hypotheses make sense, are important, and address a critical issue or need in the knowledge base. In short, the Introduction must establish that the study addresses a central issue. To the extent that the author conveys a grasp of the issues in the area and can identify the lacunae that the study is designed to fill greatly improves the quality of the report and the chances of acceptance for journal publication. By the time the readers arrive at the purpose of the study or hypotheses paragraph, they should be nodding enthusiastically and saying to themselves, "This study is really needed, it is important, it should have been done years ago, I am so glad this is being done now."

Occasionally the topic comes up about what makes a study truly important or worthwhile. This can be answered in many ways but in relation to manuscript preparation and the Introduction we are discussing; I believe the answer is "you." That is, the task is to make the case to the reader (and of course to yourself) that the study is important, interesting, and needed. We often imply that publication and communication in science are merely writing up the results or describing what has been done. Selecting important questions of course is critical but making the case that the study ought to be published, read, and added to the knowledge base is critical too. The Introduction is the chance for us as authors to do all of this. When you read a manuscript from another author, ask how strongly or well the author(s) made the case for the study.

18.3.5: Method

This section of the paper encompasses several points related to who was studied, why, and how. The section not only describes critical procedures, but also provides the rationale for methodological decisions. Subject selection, recruitment, screening, and other features ought to be covered in detail.

> **Participants and Their Selection:** Initially, the subjects or clients are described. Virtually everyone writing an article knows to do that. But in addition, provide a rationale for why this sample is included and how this is appropriate to the substantive area and question of interest.

In some cases, the sample is obviously relevant because participants have the characteristic of interest (e.g., parents accused of child abuse, adjustment, and psychological symptoms that may accompany diabetes) or are in a setting of interest (e.g., day-care center, wilderness camp). In other cases, samples are included merely because they are available. Such samples, referred to as samples of convenience, may include college students or a clinic population recruited for some other purpose than

to test the hypotheses of this study. The rationale for the sample should be provided to convey why *this* sample provides a good test of the hypotheses and whether any special features may be relevant to the conclusions. The rationale is more likely to be needed in a clinical psychology study where one might want to study something of clinical relevance (e.g., depression, trauma) and college students are used than in some other areas of psychology (e.g., social psychology) where there may be no interest in representing or talking about the applied or clinical consequences of the phenomenon. This does not mean avoid using a particular sample but rather to give some rationale for why the sample is useful, relevant, or reasonable given the goal.

Include in the description any features of the subject-selection process that might restrict the conclusions. If the sample was obtained from one or two settings (e.g., clinics), certainly note that. Also, some studies utilize participants who are part of another study. For example, one's advisor may be studying a topic and special sample (e.g., individuals who have bipolar disorder, who have a history of depression, who were special in some other way). You come along and want to do a study on something unrelated and to use data for your hypotheses. All of that is fine—even creative. In the method section, explain how the original sample was obtained (screening criteria). Later in the write-up (Discussion section) you may have to revisit the matter of whether this could restrict the external validity of the study. In any case, we want the participants described completely, to know of any inclusion or selection criteria, and to know whether there are features of the subject-selection process that could restrict the conclusions.

> **Groups Included in the Study:** The design is likely to include two or more groups that are treated in a particular fashion (e.g., experimental and control) or selected for comparison (e.g., depressed, nondepressed). The precise purpose of each group and the procedures to which they are exposed should be clarified. Control groups should not merely be labeled as such (e.g., "healthy controls") with the idea that the name is informative. It is a little better to convey precisely what the group(s) is designed to control. The author is advised to identify the critical methodological concerns and to convey how these are controlled in the design. Reviewers often criticize a study because certain control conditions were not included. After the paper is rejected by the journal, authors retort in an understandably frustrated way that the control procedure recommended by reviewers was not feasible, that the threats were not plausible anyway, and so on. Generally, the responsibility here lies with the author. The author is advised to identify the critical threats in the area and to convey how these are controlled in the design. Plausible threats that are uncontrolled deserve explicit comment to arrest the reasonable concerns of the reviewers. All of this begins in the Method section by noting what the control group is and what this is designed to control. It is not always obvious.
>
> **Assessment Devices and Procedures:** Several measures are usually included in the study. Why the *constructs* were selected for study should have been clarified in the Introduction. That is, as one weaves the rationale for the study and the background for the hypotheses, several constructs or concepts will have been mentioned.

These may include empathy, conscientiousness, mood, anger, self-control, tolerance for pain, and so on. As a general guide, *constructs* that the study is designed to test or evaluate should be reserved for the *Introduction* without mention of the measures that will be used. And the rest of this guide is that specific *measures* used to assess (operationalize) the constructs should be presented in the Method section and not in the Introduction. There are of course exceptions where studies are developing new measures or the entire basis of a study pivots on the horrible uninspired ways in which the construct has been measured in the past. Yet, the guideline usually prevails and helps clarify the significance of the study. Use the allocated pages of the Introduction to convey the strong rationales for the study and constructs you have selected. The last thing one wants is to use that limited space for procedures, again with some exceptions.

Describe the measures, especially if they are not widely familiar. Also, give information about the psychometric characteristics of the measures is often highlighted. This information relates directly to the credibility of the results. Occasionally, ambiguous, vacuous, and throw-away statements are made as one is describing the measure and its reliability or validity. For example, measures may be referred to as "reliable" or "valid" in previous research, as part of the rationale for their use. There are, of course, many different types of reliability and validity. It is important to identify those characteristics of the measure found in prior research that are relevant to the present research.

For example, high internal consistency (reliability) in a prior study may not be a strong argument for use of the measure in a longitudinal design where the author cares more about test–retest reliability. Even previous data on test–retest reliability (e.g., over 2 weeks) may not provide a sound basis for repeated testing over annual intervals. The author ought to present information to convey the suitability of the measures for the study. It is unreasonable to expect the measures to have the ideal reliability and validity data that the investigator would like to make a flawless case for use of these measures. Yet, make the case from what psychometric data there are. If data are not available,

include some analyses in the study to suggest the measure(s) behave in ways that suggest pertinent forms of reliability or validity.

Often the rationale for using a measure is that other people have used it before. That may make sense. For example, it is difficult to do an intervention study without including the Beck Depression Inventory and Hamilton Interview—these have become so standard in that literature that a departure would be seen as heresy—even though these measures and their utility are debated. Yet, in most instances, note the why you have selected the measures for the study. That can include reasons for measurement selection.

18.3.6: Results

It is important to convey why specific statistical tests were selected and how these serve the goals of the study. A useful exercise is for the investigator to read that paragraph about hypotheses and predictions from the Introduction and then immediately start reading the Results section, i.e., momentarily just skip the Method section. The results ought to speak directly to and flow from that narrative statement in the Introduction. That is, usually the hypotheses will be highlighted and then evaluated in the Results section in the same order as they were presented. If that is not advisable, convey why it makes sense to do some analyses first that are "out of order." This is not merely a style issue but rather a reflection on our thinking and on the story line we present to convey what we are doing.

Analyses often are reported in a rote fashion in which, for example, the main effects are presented and then interactions for each measure. The author presents the analyses in very much the same way as the computer output. Similarly, if several dependent measures are available, a particular set of analyses is automatically run (e.g., omnibus tests of multivariate analyses of variance followed by univariate analyses of variance for individual measures). The tests may not relate to the hypotheses, predictions, or expectations outlined at the beginning of the paper. It is important that the statistical tests be seen and presented as tools to answer questions or enlighten features of those questions and to convey this to the reader. The reader should not be able to legitimately ask, "Why was that statistical test done?"

Knowledge of statistics is critical for selecting the analyses to address the hypotheses and conditions met by the data. Yet, as important in the presentation is to convey why a given statistical test or procedure is suitable to test the hypotheses and then again what the results of that test reveal in relation to those hypotheses.

It is often useful to begin the Results by presenting basic descriptors of the data (e.g., means, standard deviations for each group or condition), so the reader has access to the numbers themselves. The main body of the Results is to test the hypotheses or to evaluate the predictions.

> Organization of the Results (subheadings) or brief statements of hypotheses before the analyses are often helpful to prompt the author to clarify how the statistical test relates to the substantive questions and to draw connections for the reader.

Think of *each paragraph* of the Results as a sandwich:

- The core or central part of the sandwich is the statistical analysis that is done to make a point or test a hypothesis
- The top slice of bread (beginning of the paragraph) is a brief statement of what we are testing (the hypothesis)
- The bottom slice of the bread (end of that same paragraph) is a brief statement that conveys what the statistics revealed in relation to that opening statement

This final statement cryptically puts into words (no numbers) what the numbers mean concretely in relation to the hypotheses. For many statistics, just presenting a tsunami of numbers, statistical tests, effect sizes, beta weights, goodness or horribleness of fit models, and so on does not obviously convey what we can conclude. Add the bottom slice so that readers can hold on to the sandwich and comfortably consume what you have done.

Several additional or ancillary analyses may be presented to elaborate the primary hypotheses. For example, one might be able to reduce the plausibility that certain biases may have accounted for group differences based on supplementary or ancillary data analyses. Ancillary analyses may be more exploratory and diffuse than tests of primary hypotheses. Manifold variables can be selected for these analyses (e.g., sex, race, height differences) that are not necessarily conceptually interesting in relation to the goals of the study. The author may wish to present data, data analyses, and findings that were unexpected; were not of initial interest; and were not the focus of the study. The rationale for these excursions and the limitations of interpretation are important to note explicitly. The excursions may generate novel hypotheses or convey something perplexing that warrants further attention in another study. From the standpoint of the reviewer and reader, the results should make clear what the main hypotheses are, how the analyses provide appropriate and pointed tests, and what conclusions can be reached as a result.

18.3.7: Discussion

The Introduction began with a statement of the need for this study and issues or lacunae in theory or research the study was designed to address. The Discussion continues

the story line by noting what we know now and how the findings address or fulfill the points noted previously. With the present findings, what puzzle piece has been added to the knowledge base, what new questions or ambiguities were raised, what other substantive areas might be relevant for this line of research, and what new studies are needed? I urge one to avoid the cliché, "this study raises more questions than it answers" but the concept behind the cliché is fine. What are a couple of the most critical questions raised by this study?

The new questions or studies referred to here are not merely those that overcome methodological limitations of the present study, but rather focus on the substantive next steps for research. As you write that, you are essentially crafting the beginning (Introduction) of the next study for yourself or another scientist to take up the matter.

More concretely, the Discussion usually includes paragraphs to provide an overview of the major findings, integration or relation of these findings to theory and prior research, limitations and ambiguities and their implications for interpretation, and future directions. These are implicit rather than formally delineated sections, and the author ought to consider the balance of attention to each topic.

Usually, the Discussion is completed within 4–5 manuscript pages. Of all paragraphs, perhaps it is useful to single out the opening one for the Discussion. Here, we provide a concise summary of the main findings.

> This paragraph may be looked at by the casual reader who goes beyond the Abstract to see a little more about what was found.

A clear paragraph right at the opening of the Discussion can be very helpful. And no need here to pluck and repeat material from the Introduction and say, "The purpose of this study was to . . . . We tested college students to see . . . ." Just go to the main findings (description), and then the rest of the discussion can focus on explanation, findings of special interest, and so on with limitations and future directions.

Methodology seems mostly relevant to the Method section. Actually, all the decisions about the study (research design, measures, statistical analyses, and more) come together in the Discussion. A tension or conflict may emerge between what the author wishes to say about the findings and their meaning versus what can be said in light of how the study was designed and evaluated. For example, as I write, I have just finished reviewing a longitudinal study on drug use in teenagers and young adults. In the Discussion, the authors talk about early factors in teen years causing later drug use. If that is what they wanted to talk about, this was the "wrong study." There was no possibility of showing a causal relation given the otherwise fine study and experimental design (observational study). But in this version of the manuscript, the conclusion and design are a misfit. This is easily corrected in this case; the authors can revise their language and talk about how some characteristic early in life is a risk factor or predictor of some later outcome. They can also do some more analyses to make less plausible other influences that might explain the findings (e.g., parental history of drinking, education of the parents, youth grades—all of which were available). The point of the example is to be sure that what can be said in the *Discussion follows from the methods, design, and analyses*. It is important to be precise about what can and cannot be asserted in light of the design and findings. A slight mismatch of interpretative statements in the Discussion and Methods is a common, albeit tacit, basis for not considering a study as well conceived and executed.

It is usually to the author's credit to examine potential limitations or sources of ambiguity of the study.

> *A candid, nondefensive appraisal of the study is very helpful.*

Here too, contextualization may be helpful because limitations of a study also are related to the body of prior research, what other studies have and have not accomplished, and whether a finding is robust across different methods of investigation. Although it is to the author's credit to acknowledge limitations of the study, there are limits on the extent to which reviewers grant a pardon for true confessions. At some point, the flaw is sufficient to preclude publication, whether or not the author acknowledges it. For example, the authors of the study might note, "A significant limitation of this study is the absence of a suitable control group. We are aware that this might limit the strength of the conclusions." Awareness here does not strengthen the demonstration itself. A huge limitation in the study is sufficiently damaging so as to preclude drawing valid inferences. It is the investigator's responsibility to convey limitations and to make the case, to the extent reasonable, that they are likely to have a minor effect, are not plausibly related to the nature of the finding, and point to issues that are logical if not important next steps for research. All studies have limitations by their very nature, so reasoning about their likely and unlikely impact on the findings is invariably relevant.

At other points, acknowledging potential limitations conveys critical understanding of the issues and guides future work. For example, in explaining the findings, the author may note that although the dependent measures are valid, there are many specific facets of the construct of interest that are not covered. Thus, the results may not extend to different facets of the construct as measured in different ways. Here too it is useful to be specific and to note precisely why other constructs and their measure might show different results. In short, be specific as to why a limitation or point might really make

a difference. This latter use of acknowledgment augments the contribution of the study and suggests concrete lines of research.

As you write up the limitations, consider this as a guide. Begin the opening paragraph in one of the usual ways in noting that there are limitations. Now note descriptively what the first limitation might be so it is very clear. Then reflect on the likelihood that the limitation really is a genuine limitation. Is it plausible or parsimonious? Are there other studies that bring to bear support for the limitation not being a problem or perhaps likely to be a problem? Then end with how that might be corrected or studied in the future. This gives each "limitation" a structure of about 3–5 sentences and draws on your expertise not merely to make a vacuous statement about a putative limitation but reflect on its likelihood and whether there might be a research question worth further study. Do this for each limitation—but only note a few limitations. (I was asked to "extensively trim" the 48-page Limitations section of my dissertation and switch to double rather than single spacing. My committee conceded that there easily were 48 pages worth of limitations but that was still too much to read.)

18.3.8: Tables, Figures, Appendices, and Other Supporting Data

There are excellent and detailed guidelines along with multiple examples of preparing tables and figures for journal articles (American Psychological Association [APA], 2010b). These comments are not a substitute for those guidelines. I mention the section here to convey broader issues. Needless to say, tables and figures are intended to provide further information and to clarify. As to the further information function, it is rarely possible to present all of the means, standard deviations, and statistical comparisons in the narrative of a Results section. Also, many of the variables (e.g., ethnicity, socioeconomic status) may not play a role in the analyses, but we want to be sure the reader knows as much as possible about the sample. The tables can be a repository of such information and also can be used to present many statistical tests and their significance. It is useful to put material in tables and to refer to it in the Results section if one can. It is often easier to see multiple tests together in a table and to make comparison across variables or groups when one can see several means, several tests, effect sizes, and p levels all together in that way.

Figures too may be a repository for information but have a stronger role in clarifying specific findings. Here ponder what picture you want to convey that is worth a thousand words. Pull out main findings or nuanced findings that are especially interesting and that can clarify.

> Some of the figures are the flow chart of subjects through the study (as illustrated later in the chapter), and others may be dictated by the data analyses (e.g., structural equation modeling).

In addition, is there some facet of the data that is worth summarizing in a chart or graphical form that gives a take home message? Try to keep figures simple so that the relation is clear. If the relation is not that clear, that too can be valuable to present in a figure, but keep what is plotted clear, simple. With the figure caption, the figure itself, and any notes at the bottom of the figure, can the reader readily glean what you intend?

The study may include an Appendix, which usually is reserved for brief material that elaborates something in the text. This might include the codes for some observational procedure, a graphic or photo of the laboratory setup and how the experimental manipulation was presented to the subjects, articles used for a meta-analysis (see APA, 2010b). The Appendix is part of the manuscript and appears with publication of the article.

Supplementary material, unlike an Appendix, does not appear in print form with the article, assuming the journal is in print form. The material is usually made available online. Supplementary material can include more detailed information about all facets of the study. Examples include details of experimental procedures or scripts provided to the subjects, more complete data analyses (e.g., if data were analyzed in multiple ways or multiple models were tested), more colored photos of brain scans, treatment or intervention manuals, and so on. Journals occasionally require the data for the study as part of the supplementary material. As authors, there is more we want to say than the limited space that most journals allow for an article. As readers, especially in our areas of expertise or specialization, we often want more details than the author provided in the printed version. Supplementary material serves both groups very well. Supplementary material is used more frequently now than ever before for several reasons:

1. The availability of online storage means that many and large files and documents readily can be linked on the Web to the article. In days of only printed materials, readers had to write to authors to obtain materials and this was onerous and not always successful as authors moved, no longer even had the materials (e.g., after 20 years), or retired from the profession or from life.
2. There is a renewed interest in the replicability of research. It is virtually impossible to repeat the procedures of a

study based on the information provided in the printed article. No one is at fault: pages are expensive to print, many articles are submitted, and authors have word and page limits. Supplementary material available online allows providing details about the study that might be of interest to only a small group of readers. Yet, replication is much more possible once one sees exactly what was said and done and what materials, tasks, and so on were presented to the participants.

3. As part of the replication, more journals are asking authors to submit the raw data and the statistical analyses along with the study. The data permit replication of the findings by allowing others to reanalyze or to consider other analyses to test the hypotheses. This is an effort to allow replication of the findings and conclusions from the data of the original study rather than replication by conducting a new study. Also, more studies now involve large databases. These databases may be made available at a central Web site and not specifically included as part of a single article.
4. Transparency has always been a value of science. Yet, with problems of replication, problems in studies where all of the information may not have been presented, and with scientific fraud in the news and a serious problem whenever it occurs, transparency (along with replication, access to data) has received heightened attention. In our everyday lives, there is the expression "too much information," to refer to conveying more than needed or wanted. (I just asked, "How are you?" and was not exactly seeking all of that really horrible information about your childhood past and "relationship issues" you are having with your significant [$p < .05$] other.) In science, the concern has been "too little information," and supplementary material is one of the strong ways to combat that. If you as an author have extensive information that may be of use to a reader or investigator replicating the study or one of those insensitive journal editors says, "Cut 3 more pages from your text," supplementary material can be the answer.

18.4: General Comments

18.4 State the primary goal of robust scientific writing

A few features in the preparation of the manuscript rise above and apply broadly to writing the individual sections:

1. Virtually all sections are guided by conveying what will be or what was done in the study (description) and the rationale for any practice (explanation). A weakness of a write-up often is the inability to identify *why* the author did this or that, i.e., the rationale was not conveyed. It is not obvious from a description why a sample, set of measures, and specific analyses were used. Merely add a sentence to convey rationales all along the way. In my view, it is worse to omit a rationale than to include one with which a reader of the manuscript might disagree. The former suggests you have not thought about what you are doing and have no real rationale; the latter conveys that you have.
2. From title to the end of the Discussion, the write-up of a study can be conceived of a story. Consider a contrasting situation. When investigators write up a study, the focus is on sections of the manuscript the way I have broken them down for presentation. Each section is written very much like adding ingredients from a recipe when baking something. When an ingredient is added, there is little reason to look back to recall what the other ingredients were or to go beyond the very next ingredient to see what is coming much later. In contrast, the "ideal" write-up is one in which there is a clear story line. That means that one can see the continuity in separate sections. The sections are not really discrete at all—the headings organize the material in obvious ways, but there ought to be continuity. I already mentioned a useful test of this. Look at the Introduction and then the Results (skip the Method section). One should be able to see, feel, and starkly note the continuity of these sections as they speak to each other, and discuss common points that clearly connect. Hypotheses are tested that were in the Introduction and it is clear which ones are tested, and how they were tested, and so on.

By story line of the manuscript, I only mean conveying what was done and how that connects to a prior section. The story line has the same characters (constructs, hypotheses), and these appear in some form throughout the plot (Results, Discussion). This is not mere repetition but connecting. Clarification of the rationale for what was done (selection of measures, data-analytic methods) across sections can help. Science writing often is noted to be very different from other writing. That is true—hyperbole, gorgeous strings of adjectives, personal digressions about one's experiences as a child leading to why a construct is of interest, mellifluous prose, and pithy universal insights usually are not welcome. On the other hand, a logical flow—themes that are clear throughout and connecting material that brings cohesion to the write up—is essential. Again, this is not a matter of style or merely style but conveying what the study accomplishes and contributes to the knowledge base.

18.5: Further Guides to Manuscript Preparation

18.5 Identify guidelines of creating a successful scientific writing

The section-by-section discussion of the content of an article is designed to convey the flow or logic of the study and the interplay of description, explanation, and contextualization. The study ought to have a thematic line throughout, and all sections ought to reflect that in a logical way. The thematic line or story line as I noted consists of the substantive issues guiding the hypotheses and decisions of the investigator (e.g., with regard to procedures and analyses) that are used to elaborate these hypotheses.

18.5.1: Questions to Guide Manuscript Preparation

A more concrete and perhaps more helpful way of aiding preparation of the manuscript is to consider our task as authors as that of answering many questions. There are questions for the authors to ask themselves or, on the other hand, questions reviewers and consumers of the research are likely to ask as they read the manuscript. These questions ought to be addressed suitably within the manuscript. Table 18.1 presents questions according to the different sections of a manuscript. The questions emphasize the descriptive information, as well as the rationale for procedures, decisions, and practices in the design and execution. The set of questions is useful as a way of checking to see that many important facets of the study have not been overlooked. As a

Table 18.1: Major Questions to Guide Journal Article Preparation

Section	Major Questions
Abstract	• What are the main purposes of the study? • Who was studied (sample, sample size, special characteristics)? • How were participants selected and assigned to conditions? • To what conditions, if any, were participants exposed? • What type of design was used? • What are the main findings and conclusions? • What are one or two specific implications or future directions of the study?
Introduction	• What is the background and context for the study? • What in current theory or research makes this study useful, important, or of interest? • What is different or special about the study in focus, methods, or design to address a need in the area? • Is the rationale clear regarding the constructs (independent and dependent variables) to be assessed? • What specifically are the purposes, predictions, or hypotheses? • Are there ancillary or exploratory goals that can be distinguished as well?
Method	• Participants ▪ Who are the participants, and how many of them are there in this study? ▪ Why was this sample selected in light of the research goals? ▪ How was this sample obtained, recruited, and selected? ▪ What are the subject and demographic characteristics of the sample (e.g., sex, age, ethnicity, race, socioeconomic status)? ▪ What, if any, inclusion and exclusion criteria were invoked, i.e., selection rules to obtain participants? ▪ How many of those subjects eligible or recruited actually were selected and participated in the study? ▪ In light of statistical power considerations, how was the sample size determined? ▪ Was informed consent solicited? How and from whom (e.g., child and parent), if special populations were used? ▪ If nonhuman animal are the participants, what protections were in place to ensure their humane care and adherence to ethical guidelines for their protection? • Design ▪ What is the design (e.g., group, true-experiment), and how does the design relate to the goals? ▪ How were participants assigned to groups or conditions? ▪ How many groups were included in the design? ▪ How are the groups similar and different? ▪ If groups are "control" groups, for what is the group intended to control? ▪ Why are these groups critical to address the questions of interest? ▪ Procedures ▪ Where was the study conducted (setting)? ▪ What measures, materials, equipment, or apparatus were used? ▪ What is the chronological sequence of events to which participants were exposed? ▪ What intervals elapsed between different aspects of the study (e.g., assessment, exposure to the manipulation, follow-up)? ▪ If assessments involved novel measures created for this study, what data can be brought to bear regarding pertinent types of reliability and validity? ▪ What checks were made to ensure that the conditions were carried out as intended? ▪ What other information does one need to know to understand how participants were treated and what conditions were provided to facilitate replication of this study?

Table 18.1 (*Continued*)

Section	Major Questions
Results	• What are the primary measures and data upon which the hypotheses or predictions depend? • What analyses are to be used, and how specifically do these address the original hypotheses and purposes? • Are the assumptions of the statistical analyses met? • If multiple tests are used, what means are provided to control error rates (increased likelihood of finding significant differences in light of using many tests)? • If more than one group is delineated (e.g., through experimental manipulation or subject selection), are they similar on variables that might otherwise explain the results (e.g., diagnosis, age)? • Are data missing due to incomplete measures (not filled out completely by the participants) or due to loss of subjects? If so, how are these handled in the data analyses? • Are there ancillary analyses that might further inform the primary analyses or exploratory analyses that might stimulate further work?
Discussion	• What are the major findings of the study? • Specifically, how do these findings add to research and support, refute, or inform current theory? • What alternative interpretations, theoretical or methodological, can be placed on the data? • What limitations or qualifiers are necessary, given methodology and design issues? • What research follows from the study to move the field forward? • Specifically, what ought to be done next (e.g., next study, career change of the author)?
More Generally	• What were the sources of support (e.g., grants, contracts) for this particular study? • If there is any real or potentially perceived conflict of interest, what might that be? • Are you or any coauthors or a funding agency likely to profit from the findings or materials (e.g., drugs, equipment) that are central to the study?

NOTE: These questions capture many of the domains that ought to be included, but they do not exhaust information that a given topic, type of research, or journal might require. Even so, the questions convey the scope of the challenge in preparing a manuscript for publication.

cautionary note, the questions alert one to the parts rather than the whole; the manuscript in its entirety or as a whole is evaluated to see how the substantive question and methodology interrelate and how decisions regarding subject selection, control conditions, measures, and data analyses relate in a coherent fashion to the guiding question.

18.5.2: Formal Guidelines for Presenting Research

In the past several years, there has been increased interest in improving the quality of research by bringing consistencies, by making procedures more transparent, and by requiring more details about the method and results. The impetus has become more salient as collaborations across disciplines have increased and science is more global. There is interest across nations in reaching common standards in relation to the openness of research, access to information, the merit-review process, and ethical issues (e.g., Suresh, 2011). Also, there has been concern about the replicability of studies and whether key procedures (e.g., are all the measures and data analyses included in the study?) are reported.

In many cases, methodologies across disciplines are shared. Perhaps the most prominent example is the randomized controlled trial, which is regarded as the gold standard for evaluating interventions. Evaluation of interventions in diverse disciplines (e.g., psychology, education, health care, pharmacology, medicine [oncology, cardiology]) usually entails investigations in which individuals are assigned randomly to various treatment and control conditions. Some of the guidelines have focused on bringing greater consistency for all the disciplines using such trials.

Several organizations and groups have developed standards for reporting research and in the process convey the need to address several facets of the study (e.g., how the sample was identified, how many started in the trial and completed the intervention, statistical power and how parameter estimates were made to calculate power, and whether participants received the intended intervention). Examples of such standards are the following:

- Consolidated Standards of Reporting Trials (CONSORT; Moher, Schulz, & Altman, 2001)
- Transparent Reporting of Evaluations with Nonexperimental Designs (TREND; Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004)
- Reporting Standards for Research in Psychology (APA, 2008)
- *Publication Manual of the APA* (Chapter 2, APA, 2010b)
- Standards for Reporting on Empirical Social Science Research in American Educational Research Association (AERA) publications (AERA, 2006)
- Meta-analytic Reporting Standards (MARS; Kepes, McDaniel, Brannick, & Banks, 2013)

The CONSORT standards, arguably the most familiar, have been adopted by hundreds of professional journals

from many disciplines and countries (see www.consort-statement.org/about-consort/supporters/consort-endorsers—journals/). The CONSORT standards and those by APA are useful to consult because they identify domains to address in preparation of a manuscript for publication and expand beyond the questions listed in Table 18.1.

In most clinical trials published in journals, authors are asked to report specifically on the flow of subjects using a special and now fairly standard flow chart. Two illustrations are provided, including a hypothetical example (Figure 18.1) and an example taken from a clinical trial of a unified cognitive behavior treatment compared to a wait list control (Figure 18.2).

Two figures are included to convey that there is not one rigid structure. Studies can vary along multiple dimensions (e.g., number of treatment conditions or "arms" of the trial, points at which participants can be lost, reasons they can be lost, and so on). The purpose is to see what happened to all potential and real participants. As one can see, the chart makes very clear what participants entered and dropped out when and who went forward to complete the study.

Beyond the flow of subjects, I mentioned the concern about increased transparency of scientific research. This applies to individual investigations but more broadly to access to many aspects of research (e.g., grant applications, descriptions of planned studies, data). In the case of federal grants (e.g., National Science Foundation, National Institutes of Health), the view is that tax payers provide support for the research and there ought to be access to the description and any materials generated from that research. The guidelines for research as well as the questions I have noted previously (Table 18.1) help convey what domains to address in the write-up of an investigation and to maximize transparency.

Figure 18.1: Sample of a CONSORT Flow Chart

Sample of a CONSORT flow chart to convey who entered the research trial and how many completed each stage

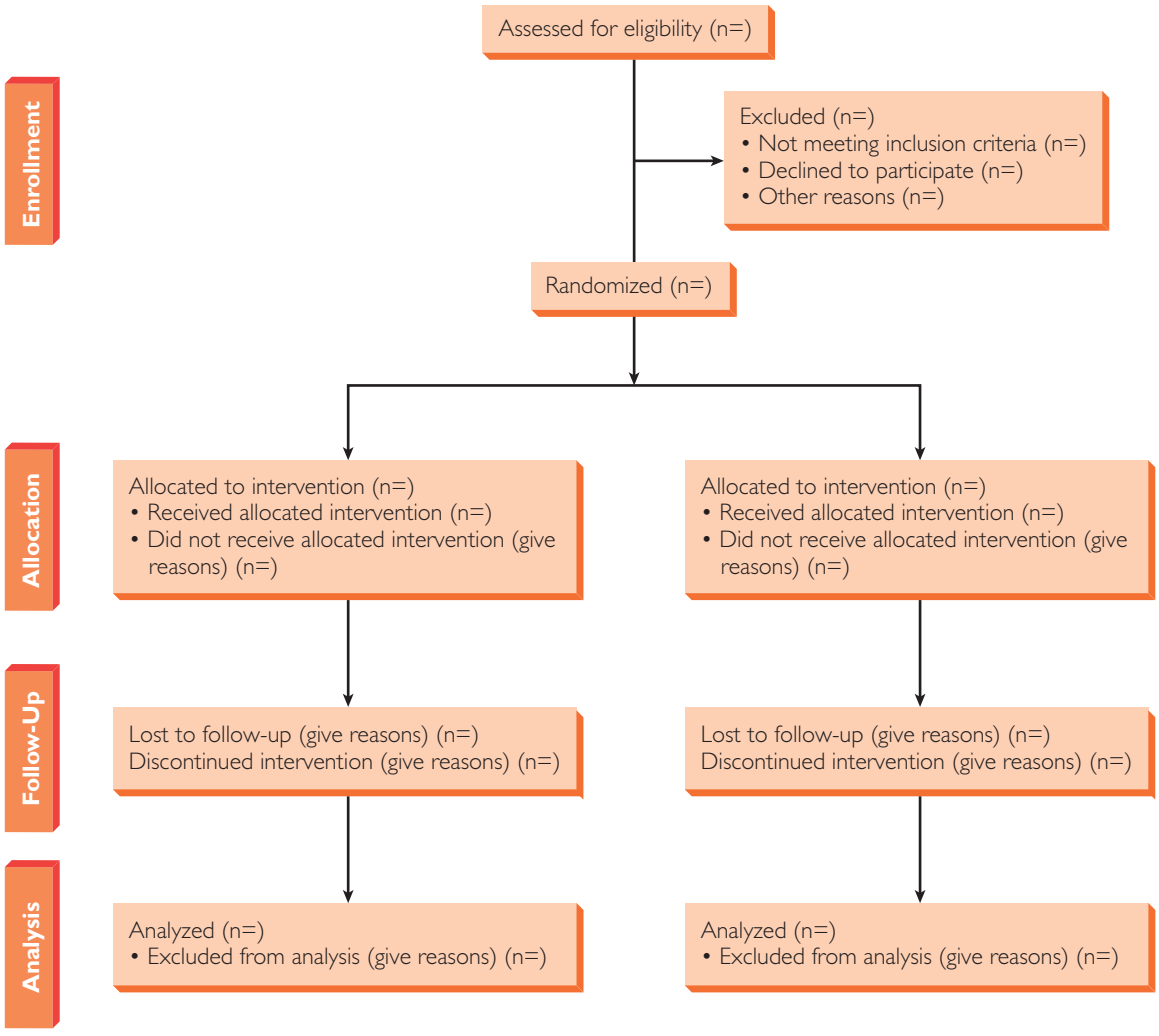
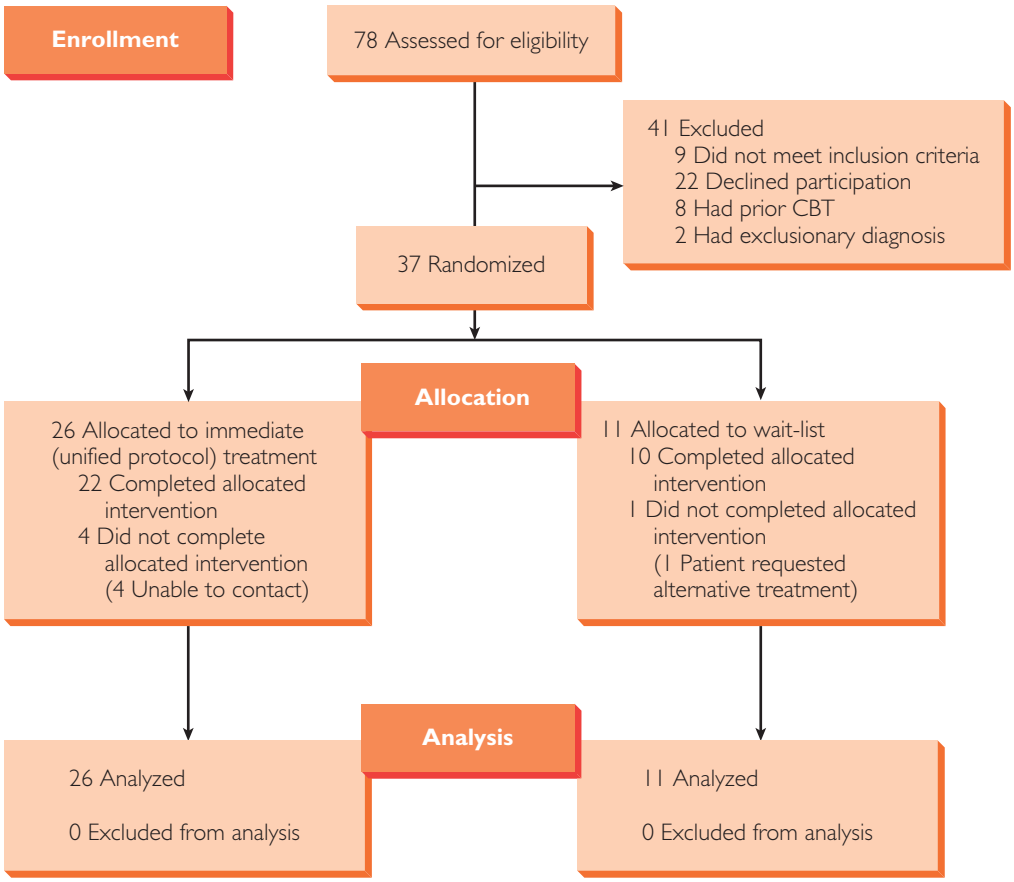


Figure 18.2: CONSORT Diagram

CONSORT diagram in a study designed to compared a unified (transdiagnostic) treatment versus waiting list control subjects



18.5.3: General Comments

Preparation of an article often is viewed as a task of describing what was done. Yet, describing a study does not automatically establish its contribution to the field, no matter how strongly the author feels that the study is a first. Also, the methodological options for studying a particular question are enormous. For example, the research question(s) could have been studied with many different samples, different constructs and measures, and data-analytic methods. The author ought to convey why a particular set of options was chosen.

In some cases, authors select options (e.g., measures, control groups) because they were used in prior research. Yet, if a key methodological decision was based solely on the argument that "others have done this in the past," that is very weak as a rationale, unless the purpose of the study is to address the value of the option as a goal of the study. Also, it may be that new evidence has emerged that makes the past practice more questionable. Over time, the standards and permissible methods may change, and measures and controls once viewed as innovative or even acceptable are viewed as mundane or flawed.

In general, it is beneficial to the author and to the scientific community more generally to convey the thought processes underlying methodological and design decisions. This information will greatly influence the extent to which the research effort is appreciated and viewed as enhancing knowledge. The author is not advised to write a persuasive appeal about how important the study is and how this or that way was the best way to study the phenomenon. Yet, it is useful to convey that decisions were thoughtful and that they represent reasonable choices among the alternatives for answering the questions that guide the study. The contextual issues are no less important. As authors, we often expect the brilliance of the study to shine through and to be self-evident. Yet, the contribution of a study is a judgment call. From the perspective of the author, it is advantageous to be very clear of how and where the study fits in the literature, what it adds, and what questions and research the study prompts.

18.6: Selecting a Journal

18.6 Recognize the importance of selecting the appropriate journal for scientific publication

I discuss journal selection here because it logically follows in the sequence of steps we have been covering. We complete a study, prepare the write up, and submit the article for publication. Selecting a journal is part of the process of communicating one's results and completes this part of the process.

Preparation of the manuscript is logically before selecting a journal and submitting the journal for publication. However, investigators occasionally have the journal or a couple of journals in mind before the manuscript is prepared. Journals have different emphases and research with:

- Particular sorts of foci (e.g., theory, application)
- Samples (e.g., nonhuman animals, college students, community samples)
- Settings (laboratory, field)
- Research designs (cross-sectional, longitudinal, experimental, observational)

Consequently, it is not odd for the investigator to plan/hope that a study when completed will be appropriate for a journal he or she targeted well before preparing the manuscript for publication.

18.6.1: What Journal Outlets Are Available?

Several resources convey possible journal outlets available for research in the behavioral and social sciences, and the resources and potential relevance to your study are easily obtained from the Web (Gunther, 2011; Thursby, 2011; Thomson Reuters, 2011). These sources can be searched by topic and key words in relation to how you view your study (e.g., clinical psychology, my likely candidacy for a Nobel prize). There are many professional organizations and smaller societies within psychology that have their own publications.

> The two major professional organizations whose journal programs are widely recognized and emulated are American Psychological Association (APA) and the Association for Psychological Science (APS).

For example, the APA lists over 80 journals in psychology within the English language and aids author in selecting an appropriate outlet (APA, 2014; www.apa.org/pubs/journals). Also, on the Internet there are excellent sites that search for journals in given area (e.g., clinical psychology) and provide (at this writing) access to over 2,000 journals (www.psycline.org/journals/psycline.html). These sources provide information that includes the editorial policy, content area or domain, type of paper that will be considered (e.g., investigations, literature reviews, case studies), guidelines for manuscript preparation for each of the journals, and tables of contents of current and past issues. Journals continue to proliferate, especially with the development of online and open access journals without any printed version.

I have emphasized journals in the English language. Psychology is an active discipline internationally, and psychological associations in many countries and regions (e.g., European Union, Scandinavia, Asia) have many excellent journals as well. Many such journals publish articles in English and also may include a translation as well. In the United States, but internationally as well, the number of journals continues to grow. So there are innumerable options to find a match for your study. It would be a great idea if there were a match making site that puts manuscripts and journals together—kind of like one of those dating-find-your-mate Web sites—maybe something like www.pleasepublishmywork.com.

18.6.2: Some Criteria for Choosing among the Many Options

How to select among the journals has several potential criteria and considerations. In my own case, sometimes I prefer to see the final or almost final write-up to consider what journals might be reasonable outlets for the article. (I like journals that are desperate for articles or those with some blood relatives on the editorial board.) On other occasions, there is a special audience I wish to reach (e.g., mental health researchers rather than only psychologists, child treatment researchers rather than treatment researchers more generally, and so on as they relate to my study). This latter criterion is based on who is likely to read or subscribe to the journal. Also relevant, journals vary markedly in their readership and subscription base. Some journals have subscribers who can vary from relatively few (e.g., 200–600) to several thousands, are in accessible through few libraries, or are omitted from easily accessed databases the library has purchased for use. Fortunately, most professional journals have their Abstracts included in databases that can be accessed from the Web. This makes even the most obscure study accessible.

Let me begin with two easy guidelines for considering where to submit a manuscript for publication. First, what journals are being cited in your manuscript? As you complete the write-up, peek one more time at the Reference section.

Are there some journals cited frequently?

Perhaps articles were cited in the Introduction or Discussion of the manuscript that comes from the same journal. If the same journal comes up a few times in the References, this suggests that your manuscript may be

speaking to the audience (researchers, readers) of that journal. It also may indicate that the journal has precedent for publishing on the topic. So one consideration may be to see what journals are being relied on to make the case for the study.

Second and related, consider the above point but at a higher level of abstraction. Perhaps no one journal dominates the Reference section, but a topic or area of study does. For example, perhaps many of the journals focus on violence, victimization, aggression, or emotion regulation or even more broadly and abstractly, cognition or affect.

> Consequently, now ask yourself, "Is there an emergent theme that includes many of the journals in the Reference section even though no one journal dominates that section?"

This might make the selection of journal focus on any one of the journals you cited that is in keeping with that theme, even though no particular journal was cited often. Of course, if you are working with an advisor or someone who has a program of studies that has been going on for a while, the options and likely publication outlets will be much clearer from the outlets used in the past.

18.6.3: Additional Criteria for Consideration

Many additional criteria are invoked to select a journal to which one will submit a manuscript, including the relevance of the journal in relation to the topic, the prestige value of the journal in an implicit hierarchy of journals in the field, the likelihood of acceptance, the breadth and number of readers or subscribers, and the discipline and audience one wishes to reach (e.g., psychology, psychiatry, medicine, social work, health, education). As for the prestige value, clearly some journals are regarded as more selective than others. For example, some of the APA journals are premier journal outlets in their respective areas. In clinical psychology for example, two journals that are widely recognized as the premier journals are:

- *Journal of Abnormal Psychology*
- *Journal of Consulting and Clinical Psychology*

For APS, there is one clinical journal (*Clinical Psychological Science*), that too is very selective in what is accepted for publication.[3] Yet, journals from other organizations, journals not sponsored by an organization, and journals from other professions or disciplines can be as or more highly regarded. Indeed, in some areas (e.g., social, cognitive, or clinical neuroscience), some of the most discriminating and selective publication outlets in which psychological research occasionally is published are not psychology journals (*Science*, *Nature*, *Neuroscience*). One can identify the best outlets by familiarity with the literature (e.g., where do the best studies seem to be published) and by chatting with colleagues. There is such increased specialization in research that the "best" home of one's article might well be in a journal that definitely reaches the audience doing the same general kind of work.

Reputation of a journal and the quality of its articles usually are well known. There has been an enduring interest in having more objective measures and they are available. The impact of a journal is primary among these measures (Web of Science, 2011) and includes the extent to which articles in a journal are cited by others. Journals with articles that are heavily cited are those with much higher impact. Information is available for journals in virtually all areas of science. Within the social sciences alone, over 2,000 journals are covered.[4]

Whatever journal is selected, identify one that uses peer-review. This of course refers to the fact that manuscripts undergo a process where anonymous reviewers evaluate the manuscript to decide on its suitability for publication. There is an enormous proliferation of journals now, many online journals also referred to as electronic journals or e-journals. Some of these such as those published by the Public Library of Science (known as PLoS) undergo peer review and appear in databases (e.g., PubMed maintained by the National Library of Medicine) that select based on the quality and reputation of the journal. Another is *eLife*, an electronic journal that encompasses the life sciences but spills into areas of psychological science too (e.g., neuroscience, health) (see http://elife.elifesciences.org/about-the-journal). A new online journal (*Archives of Scientific Psychology*) has emerged from the APA (Cooper & VandenBos, 2013, www.apa.org/pubs/journals/arc/index.aspx) and publishes articles from all areas of psychology. This is a free and open access journal but with strong requirements for reporting (using standards I mentioned previously) and is peer-reviewed. These journals adopt a rigorous review process and are recognized to be discriminating in what they select and publish. Despite examples noted here, many online journals are not so easily identified as peer-reviewed and caution is needed in their selection.

Some journals, print or online, are not very selective and, indeed, have to hustle (e.g., invite, accept many) articles so that they can fill their pages. The more obscure and low impact journals may actually be in a little trouble in accepting enough papers. A few journals in psychology charge authors money for publishing their papers. So when one's paper is accepted, the author is charged on the basis of how many journal pages the article will require. These outlets do not necessarily take all submissions, but they often take most. Within psychology, career advice is to focus on peer-reviewed and well-regarded journals, leaving aside other issues.

Knowledge of the area of research and contact with one's colleagues can readily identify the ideal outlets for one's research.

As a guide, one's research is evaluated in part by the company it keeps, i.e., where it is published and the standards of that journals.

Most journals are in print (hard copy) and electronic form but many are only Web based and others are moving toward that format. This is not the place to discuss that topic except to note that often publication on the Web can be much faster (less delay in review of the manuscript and acceptance of the manuscript) than is publication in a printed journal. There are still dynamic changes in how journals will be published and disseminated and print versions probably are on borrowed time.

The central issue for one's career is the extent to which the publication outlet is well regarded by one's peers and the care with which manuscripts are reviewed before they are accepted and published.

Electronic versus printed journal format is not as critical as the quality of the publication. If publication in the journal requires little or no peer review, if most manuscripts are accepted, and if manuscripts are accepted largely as they are (without revision), quality of the research and the value of the publication to one's career may be commensurately reduced.

18.7: Manuscript Submission and Review

18.7 Detail the scientific publication submission and review processes

Alas, after very careful deliberation, a 30-minute discussion with your coauthor(s), and 2 hours at a Ouija board with your significant other ($p < .05$), you finally select a journal and are ready to submit your manuscript for publication. Before you do, consult the Instructions to Authors written in the journal to make sure that you submit the manuscript correctly. Usually manuscripts are submitted through a journal portal, i.e., electronically in which the manuscript file and a letter of submission are uploaded to the journal Web site. This is the usual way of submitting a manuscript whether the journal is an online-only journal or a more familiar print journal. In some cases, you may be required to include sentences or paragraphs in the letter you submit that say this study is not being considered elsewhere in another journal, has not been published before, that you will give the copyright to the publisher if the manuscript is accepted, and that your study met appropriate human rights protection guidelines. Processing of the manuscript could be delayed if your letter does not meet the guidelines provided in the journal.

18.7.1: Overview of the Journal Review Process

Once the manuscript is submitted, the journal editor usually sends the electronic file to two or more reviewers who are selected because of their knowledge and special expertise in the area of the study or because of familiarity with selected features of the study (e.g., novel methods of data analyses). Sometimes you may even be asked to identify possible reviewers. These are mere recommendations, and editors may or may not use the information.

Reviewers may be selected from the names of authors whose articles you included in your Introduction.

Some reviewers are consulting editors who review often for the journal and presumably have a perspective of the type and quality of papers the journal typically publishes; other reviewers are ad hoc reviewers and are selected less regularly than consulting editors.

Reviewers are asked to evaluate the manuscript critically and to examine whether or the extent to which:

- The question(s) is important for the field
- The design and methodology are appropriate to the question
- The results are suitably analyzed
- The interpretations follow from the design and findings
- The knowledge yield contributes in an incremental way to what is known already

All the threats to experimental validity (internal, external, construct, data evaluation) and potential sources of bias are quite fair game as the foci of the reviewers' comments. While the specific features of the study are critically important, they are the "parts." The "gestalt" or whole of the manuscript is about the clarity of the story line and is arguably as or more important than the parts.

The task of the reviewers is to evaluate the manuscript in light of all of the considerations I have mentioned and indeed any others that reviewers wish. Typically, reviewers are asked to give a summary recommendation (e.g., reject or accept the manuscript). All recommendations to an editor are advisory and not binding in any way. At the same time, the editor sought experts and usually follows their recommendations if there is a consensus that emerges. Yet reviewers too must make the case for their comments. Many editors do not really read the manuscripts and default to the reviewers.

When reviewers submit their comments, usually they provide secret comments to the editor. These are comments we as authors do not see. Here the reviewer might give a few summary comments, a more candid opinion of the work, or convey dilemmas in making a recommendation. Reviewers also have a section to provide comments to the author. These comments convey issues, concerns, and

recommendations to the author and provide the underpinnings of the recommendation the reviewer gave.

Once the paper is reviewed, the editor evaluates the manuscript and the comments of the reviewers. In some cases, the editor may provide his or her own independent review of the paper; in other cases, he or she may not review the paper at all but defer to the comments and recommendations of the reviewers. The editor writes the author and notes the editorial decision. Usually, one of three decisions is reached:

- The manuscript is accepted pending a number of revisions that address points of concern in the reviewers' comments
- The manuscript is rejected and will not be considered further by the journal
- The manuscript is rejected, but the author is invited to resubmit an extensively revised version of the paper for reconsideration

The ***accept*** decision usually means that the overall study was judged to provide important information and was well done. However, reviewers and the editor may have identified several points for further clarification and analysis. The author is asked to revise the paper to address these points. The revised paper would be accepted for publication.

The ***reject*** decision means that the reviewers and/or editor considered the paper to include flaws in conception, design, or execution or that the research problem, focus, or question did not address a very important issue. For the journals with high rejection rates, papers usually are not rejected because they are flagrantly flawed in design. Critical potential threats to validity presumably have been well handled. Rather, the importance of the study, the suitability of the methods for the questions, and specific methodological and design decisions conspire to serve as the basis for the decision.

> The *reject-resubmit decision* may be used if several issues emerged that raise questions about the research and the design. In a sense, the study may be viewed as basically sound and important but many significant questions preclude definitive evaluation.

18.7.2: More Information on Overview of the Journal Review Process

The author may be invited to prepare an extensively revised version that includes further procedural details, additional data analyses, and clarification of many decision points pivotal to the findings and conclusions. The revised manuscript may be re-entered into the review process and be evaluated again.

Of the three letters, clearly a rejection letter is the most commonly received. Authors and perhaps new authors in particular are not sufficiently prepared for this feature of the journal publication business.[5] Journals often publish their rejection rates, i.e., proportion of submitted manuscripts that are rejected, and this figure can be quite high (e.g., 70–90%). Often the prestige value of the journal is in part based on the high rejection rate. Yet, the rate is ambiguous at best because of self-screening among potential authors. For very prestigious publication outlets (e.g., *Psychological Review, Science*) where psychological papers are published, the rejection rates area obscured by the self-screening that most authors are not likely to even try that outlet if they have a contribution that falls within the topic and format domain. Rejection rates across journals are not directly comparable. Even so, the rates give the would-be author the approximate odds of authors entering the fray.[6]

Although beyond our purpose, the review process deserves passing comment. The entire process of manuscript submission, review, and publication has been heavily lamented, debated, and criticized. The peer-review process has a long history as an effort of quality control over the content and standards of what is published (Alvarez, Fernández, Conroy, & Martínez, 2008; Spier, 2002). The alternatives to peer review (e.g., no review, judgment by one person such as the editor) have their own liabilities. Many journals invoke procedures where the identity of the authors and the reviewers is masked, i.e., names are not included on the manuscript sent to reviewers or the reviews sent to authors. The goal is to try to limit some of the human factors that can operate about responses to a person, name, or other facet and to allow reviewers to be candid in their evaluations without worrying about facing the colleague who will never speak to them again. The peer-review system is far from perfect. The imperfections and biases of peer review, the lack of agreement between reviewers of a given paper, the influence of variables (e.g., prestige value of the author's institution, number of citations of one's prior work within the manuscript) on decisions of reviewers, and the control that reviewers and editors exert over authors have been vigorously discussed for decades (e.g., Cicchetti, 1991; Lindsay, 1988; Smith, 2006).

Understanding the review process can be aided by underscoring the one salient characteristic that authors, reviewers, and editors share, to wit, they are all human. This means that they (we) vary widely in:

- Skills
- Expertise
- Perspectives
- Sensitivities
- Motives
- Abilities to communicate

Science is an enterprise of people and hence cannot be divorced from subjectivity and judgment. In noting subjectivity in the manuscript review and evaluation process, there is a false implication of arbitrariness and fiat. Quality research often rises to the top, and opinions of quality over time are not idiosyncratic. Think of the peer-review process as the home-plate umpire in a baseball game. Any given call (e.g., strike) may be incorrect, arguable, and misguided. And any given pitcher or batter suffers unfairly as a result of that call. As reviewers (the umpires) make the call on your manuscript (rejection, you strike out), you too may have that occasional bad call. But over time, it is unlikely that all manuscripts an author submits receive a misguided call. Pitchers and batters earn their reputations by seeing how they perform over time, across many umpires, and many games. One looks for patterns to emerge, and this can be seen in the publication record of an active researcher.

18.7.3: You Receive the Reviews

Alas, the editorial process is completed (typically within 3 months after manuscript submission) and the reviews are in. You receive an e-mail from the editor noting whether the paper is accepted for publication and if not whether it might be if suitably revised. It is possible that the e-mail will say the manuscript is accepted as is (no further changes) and praise you for your brilliance. If this occurs, it is the middle of the night and you are dreaming. Remain in this wonderfully pleasant state as long as you can. When you awake, your spouse or partner reads the printed version of the real e-mail and you read one of the three decisions noted previously.

> If the manuscript is accepted, usually some changes are needed. These do not raise problems. More often than not, the manuscript is rejected.

There are individual differences in how one reacts to this decision. Typically, one feels at least one of these:

- Miffed
- Misunderstood
- Frustrated
- Angry at the reviewers

Usually one has only the e-mail comments and has limited avenues (e.g., scrutiny of the phrasing and language) for trying to identify who could have possibly rejected the manuscript. If a hard (printed) version of the reviews was sent, one can scrutinize the font style, key words, possible DNA remnants of the reviewers' comments sheets, and molecules on the pages that might reveal pollutants associated with a particular city in the country. To handle a rejection verdict, some authors select one of the very effective psychotherapies or medications for depression; others use coping strategies (e.g., anger management training, stress inoculation therapy) or other procedures (e.g., acupuncture, mineral baths, vegan enemas). (I myself use all these routinely with their order counterbalanced to help redress multiple-treatment interference.)

The task is to publish one's work. Consequently, it is useful and important to take from the reviews all one can to revise the manuscript. Maladaptive cognitions can harm the process. For example, when reading a review, the author might say, "The reviewer misunderstood what I did or did not read this or that critical part." These claims may be true, but the onus is always on the author to make the study, its rationale and procedures, patently clear.

> *A misunderstanding by a reviewer is likely to serve as a preview of the reactions of many other readers of the article.*

Indeed, most readers may not read with the care and scrutiny of the reviewers. If the author feels a rejected manuscript can be revised to address the key concerns, by all means write to the editor and explain this in detail and without righteous indignation and affect.

Authors often are frustrated at the reactions of reviewers. In reading the reactions of reviewers, the authors usually recognize and acknowledge the value of providing more details (e.g., further information about the participants or procedures). However, when the requests pertain to explanation and contextualization, authors are more likely to be baffled or defensive. This reaction may be reasonable because much less attention is given to these facets in graduate training. Also, reviewers' comments and editorial decision letters may not be explicit about the need for explanation and contextualization. For example, some of the more general reactions of reviewers are often reflected in comments such as: Nothing in the manuscript is new, "I fail to see the importance of the study," "This study has already been done in a much better way by others."[7] In fact, the characterizations may be true. Authors (e.g., me) often feel like they are victims of reviewers who wore sleep masks when they read the manuscript, did not grasp key points, and have had little exposure to, let alone mastery of, the pertinent literature. Occasionally two or more of these are true.

As often as not, it is the reviewers who might more appropriately give the victim speech. The author has not made the connections among the extant literature and this study and integrated the substantive, methodological, and data-analytic features in a cohesive and thematic way. Reviewers' comments and less than extravagant praise often reflect the extent to which the author has failed to contextualize the study to mitigate these reactions. The lesson for preparing and evaluating research reports is clear. Describing a study does not establish its contribution to the field, no matter how strongly the author feels that the study is a first.

Let us assume that the manuscript was rejected with an invitation to resubmit. As a rule, I try to incorporate as many of the reviewers' and editor's recommendations as possible. My view is that the reviewer may be idiosyncratic, but more likely represents a constituency that might read the article. If

I can address several or all issues, clarify procedures that I thought were already perfectly clear, and elaborate a rationale or two, it is advisable to do so. Free advice from reviewers can and ought to be used to one's advantage.

There are likely to be aspects of the reviews one cannot address. Perhaps reviewers provide conflicting recommendations or a manuscript page limit precludes addressing or elaborating a particular point. Even more importantly, perhaps as an author one strongly disagrees with the point. Mention these in the letter to the editor that accompanies the revised manuscript. Explain what revisions were or were not made and why. If there are large revisions that alter the text (few sentences), methods, or data analyses, help the editor by noting where the change can be found in the manuscript and even submit an extra copy of the manuscript in which the changes are tracked in some editing/word processing system.

The investigator may receive a rejection letter and decide simply to submit the manuscript as is to another journal. I believe this is generally unwise. If there are fairly detailed reviews, it is to the author's advantage to incorporate key points and often not-so-key points, even if the manuscript is to go to another journal. I have often seen the same manuscript (not mine) rejected from two different journals in which there were no changes after the first rejection. The authors could have greatly improved the likelihood of publication in the second journal but were a bit stubborn about making any revisions. Even if the manuscript were to be accepted as is in the second journal, it is still likely the author missed an opportunity to make improvements after the first set of reviews was provided.

In general, try to take all of the recommendations and criticisms from the reviews and convert them to facets that can improve the manuscript. Obstacles to this process may stem from our natural defensive reactions as authors or a negativity bias and the occasional brutish way in which reviewers convey cogent points. (I remember being highly offended the first two or three times reviewers noted such comments, "the author [me] would not recognize a hypothesis if it fell on his lap" and "the design of this study raises very important issues, such as whether it is too late for the author [me] to consider a career change." I have come to refer to all of this as the *pier*-review process to underscore how often reviewers have made me want to jump off one.)

It is worthwhile and highly rewarding to publish one's research. The process takes time and persistence. Also, contact with others through the review process can greatly improve one's work. In my own case, reading the reviews occasionally has stimulated next studies that I carried out. In one case, I befriended a person who was a reviewer of my work earlier in my career. Over time and from following his work, it was very clear that he was behind an influential review, although his identity had been masked. Years later over dinner, I mentioned his review in a distant past, the study it generated, and the very interesting results and, of course, expressed my gratitude. His suggestion actually led to a few studies. (His review of my manuscript was not entirely positive, which probably is the main reason I hid in bathroom of the restaurant until he paid the check for dinner.) The lesson is more than getting one's manuscript published. Reviews can be very educational, and it is useful to let the comments sit for a while until the rage over rejection subsides.

18.7.4: General Comments

I have focused my comments on publication of empirical research in journals. I gave this emphasis because writing up studies whether for journal publication or other venues (e.g., senior theses, venues) shares common characteristics. Also, journal publication is the commerce of psychological science—that is what researchers do.

All of that said, there are other types of papers than empirical studies. Many authors write review papers such as meta-analyses to evaluate progress in an area and to ask questions of the overall literature. Meta-analyses remain a common way of reviewing the body of evidence, as I have commented earlier in the book.

Also, one can disseminate findings from a study by presentation at various professional meetings and conferences. Here the research may be presented and delivered via a power point presentation. Alternatively, there are poster sessions where the study and results are summarized on a large board and passersby look at the summary and chat with the investigator about what was done. In short, there are other options for communication of one's findings beyond the primary one I have discussed in this chapter (see Prinstein, 2013). With these different formats for communication of one's results, the broad points and guidelines remain the same.

- Make a strong case for why the study is needed or what it will contribute (e.g., beyond platitudes that prior studies had this or that methodological limitation unless you can show that is really important)
- Convey the rationale for the study and its many parts (e.g., why this sample, why these measures, why these data analyses)
- Connect the sections so that the story line is clear (e.g., the organization and of the Results and Discussion sections should be closely connected to what was stated in the Introduction)
- Convey the limitations (e.g., any threats to validity not well addressed? Also note why or why not a particular limitation is not really a limitation or if it is what future research should do about it)
- Highlight the next steps for research (e.g., what is the next study or line of work needed to make progress) and here too why are those next steps important

Summary and Conclusions: Communication of Research Findings

Communication of results of research represents a complex process involving many issues beyond methodology and research design. In preparing a written report, diverse abilities are taxed beginning with the author's skills in identifying and selecting critical substantive questions and culminating with skills in communicating the results. Methodology and design play major roles throughout the processes of planning, conducting, and communicating research results.

Three interrelated tasks are involved in preparing a manuscript for journal publication. These were described as description, explanation, and contextualization of the study. The writing we are routinely taught in science focuses on description, but the other portions are central as well and determine whether a study not only appears to be important but also in fact actually is. Recommendations were made in what to address and how to incorporate description, explanation, and contextualization within the different sections of a manuscript (e.g., Introduction, Method). In addition, questions were provided to direct the researcher to the types of issues reviewers are likely to ask about a manuscript.

In preparing the manuscript, the author invariably wishes to make a statement (conclusion). The strength of that conclusion is based on the extent to which the study addresses issues highlighted in prior chapters. It is important for the author to convey the focus and goals of the study clearly and concisely. The design decisions and the rationale for these decisions, when presented clearly, greatly augment the manuscript. Also, I underscored the importance of a clear story line that refers to the theme that unites the various sections of a manuscript. To the extent one can, it is useful to convey the continuity of the main sections (e.g., Introduction, Method, etc.) where each section looks back as it were to reflect the organization and main themes of that section.

Also discussed were criteria for selecting a journal as an outlet for one's work. There are several criteria I noted. One is looking for a journal outlet that will reach the intended outlet and that is regarded as an excellent peer-reviewed journal. There are multiple other considerations that might be invoked as well (e.g., speed of publication, audience that will be reached). Within a given area of research (e.g., clinical neuroscience, research on psychiatric disorders and their etiology, intervention research), there is reasonable consensus of the desired publication outlets. If in doubt, one's colleagues are usually a great source of information.

Critical Thinking Questions

1. What is meant by contextualization, and why is that critical when making the case for doing or reporting an investigation?
2. Telling the reader in a report or reading in a report of others that this is the "first study" or that the study "overcomes some methodological gap of a prior study" is not automatically important. Why?
3. What would be two or so general guidelines you would advise others in preparing a written report of their study? Why are these important?

Chapter 18 Quiz: Communication of Research Findings

Chapter 19
Methodology: Constantly Evolving along with Advances in Science

Learning Objectives

19.1 Analyze the complementary roles of evolving scientific methodologies and advancing scientific knowledge

19.2 Describe research design

19.3 Evaluate the importance of multiple scientific research methodologies

19.4 Review the guidelines to effective statistical methodologies

In this chapter, we will discuss the key messages that one might draw from what research methodology is and does.[1] This is especially important because methodology in psychology often is taught in a way that emphasizes specific practices or ingredients, kind of like a cookbook. That is, there are special ingredients and they can be combined in various ways to produce a study:

- Add one or two hypotheses
- Select lots of fresh subjects, obtain informed consent, and mix thoroughly (but randomly)
- Stir in three or more measures or assessment devices
- Collect, score, and enter the data from the completed measures
- Allow to cook for a few nanoseconds while statistical analyses are processed
- Generate *F* or *t* tests, maybe a regression analysis, mediation test, or output from a vast array of other statistical tests (some might even add "beat the ingredients" until they obtain statistical significance)
- Describe the study in a clear but cryptic structured style (APA, 2010b, publication manual)
- Submit the manuscript to a journal relevant to the topic of your study
- Allow to sit for 3 or so months (the review process) and maybe 12 more months (actual publication)
- Send a pdf file of the published article to relatives and hope they do not say, "no thanks" like mine did the last time I did this

Additional Information on Methodology

No doubt the art and science of methodology make cooking a reasonable metaphor, but not very helpful. Clearly, there are ingredients that can characterize sound studies, but it is better to move away from the ingredients a bit to understand the task and goals of methodology more broadly.

We have considered methodology at the course at multiple levels:

1. Methodology at a very general level is a way of thinking and problem solving. That includes a commitment to scientific evidence and a way of describing and explaining the world.

   Concepts such as parsimony and plausible rival hypotheses are central to the thought processes underlying methodology.

   Related, we are concerned about how one can move from a general idea to a hypothesis and then to an empirical test of that hypothesis.

   Understanding and developing testable hypotheses and operational definitions also are key concepts at this general level.

2. The thinking and problem-solving facets of methodology direct attentions to several concerns or issues when designing a study or evaluating a study completed by someone else. All the threats to validity (internal,

external, construct, and data evaluation) are at this more specific level. Ideally, as one evaluates a study to be conducted or already completed, one's mind would go to each of the threats and identify where attention is needed. Consider this part akin to medical diagnosis—you are looking at the patient and want to discover health, wellness, but also illness. What to test, perhaps we begin by taking a blood sample (routine blood work), listening to breathing and heartbeat, and measuring blood pressure. Nothing here is sophisticated and just the very basics.

In methodology, threats to validity are the basic diagnostic tools to plan and evaluate the study. If you have not memorized the threats to validity, use the laminated list of threats that you may now have made for your wallet. In the study you are designing or in the manuscript of the study you are reading, is each of the threats well managed? If not controlled, how much of a problem is it? Remember a threat to validity is only a real threat if that factor can plausibly explain the results or jeopardize the investigator's conclusions. The nice feature is that for most of the threats to validity, we can plan ahead to decrease the likelihood that they will interfere with drawing valid inferences. And as we are reading the study someone else has completed, we know that several threats are often controlled at once (e.g., by random assignment) and several threats can be jeopardized by single problems that emerge (e.g., weak power, attrition).

3. There are the specific practices that form methodology. These are the concrete parts that are usually taught. These are important and include a pile of "how to" and "what to do":

   For example, practices include how to:

   - Randomly assign participants to conditions
   - Calculate power to decide how large the sample should be
   - Minimize and control variability (error) by monitoring implementation of the procedures, selecting reliable measures, and so on

   And practices that include the what:

   - Select among control groups depending on the question and alternative hypotheses that one ought to make implausible
   - Select conditions (among experimental and control groups) that provide the strongest test of one's hypotheses
   - Decide what variables might need to be controlled either by matching subjects or by addressing these statistically when evaluating the results
   - Select an antidepressant, evidence-based therapy, or meditation to cope with the comments from reviewers if you have submitted a manuscript for publication

Methodology at all three levels is critical to develop, design, and implement a well-conducted study. Yes, there is no substitute for a creative idea that serves as the basis for the study, and we discussed sources of ideas to help reach that point.

> Providing a suitable test that is the best way to provide support for that hypothesis requires understanding the challenges, considering the threats to validity and potential sources of bias in mind, and then deciding what practices will be optimal and feasible.

In short, one message of the text is the importance of methodology and its different levels, including but well beyond a plethora of tricks and concrete practices.

19.1: The Dynamic Nature of Methodology

19.1 Analyze the complementary roles of evolving scientific methodologies and advancing scientific knowledge

The natural, biological, and social sciences are making breathtaking advances. Only a minute portion makes the news, and of course those are selected precisely because of their likely mass appeal (e.g., revelations about planets outside of our solar system that might be habitable, new diagnostic tests or treatments for an otherwise incurable disorder, more good news on how wonderful chocolate is for physical health, and bad news that something we thought helps [diet drinks] actually have the opposite effect by increasing our calorie consumption). In psychology, there are truly remarkable findings on questions of great interest to the public at large, and they occasionally make the news too. The list of remarkable findings is virtually endless, but an unsystematic sample conveys the point. As examples, from randomized trials:

- We know how to improve children's IQ
- We know that reading literary fiction can change the brain and facilitate other learning
- Enduring stress can damage the immune system, increase aging, and shorten a person's life (e.g., Glaser & Kiecolt-Glaser, 2005; Kidd & Castano, 2013; Protzko, Aronson, & Blair, 2013)

It is natural to focus on the substantive findings that are intriguing and address concerns related to our daily lives. Yet, quietly behind the scenes in all of this are advances in methodology. I say quietly because there are not too many headlines on cable and network news that are like, "New Way to Randomize Subjects" or "What You Need to Know about Threats to External Validity." Sad—many of us scan the news looking for these the way

scientists who scan the cosmos are hoping for radio signals from beings from other worlds. I am not making up the fact that advances in methodology (e.g., assessment or data evaluation) are central to the many substantive advances in science (e.g., identifying the Higgs-Boson particle in physics, understanding climate change), including areas that are core parts of clinical psychology (e.g., neuroimaging, genetics, observational designs) (see Fienberg, 2014; Greenwald, 2012).

Methodology is constantly changing and developing, and these changes underlie the further advances. Consider illustrations merely to support the point rather than to comprehensively cover these advances and changes in any one area. I mentioned at the outset of the text that methodology usefully is conceived as having five components:

- Research design
- Assessment
- Data evaluation
- Ethical issues
- Communication of research findings

Each of these is changing and evolving in many ways, although they are not in public view beyond the substantive findings they foster. Consider some of the advances briefly.

19.2: Research Design

19.2 Describe research design

Research design refers to the arrangements of the study.

We discussed having broad categories (e.g., true-experiments, observational designs, single-case designs) as well as specific designs within them (e.g., randomized controlled trials [RCTs], case-control designs). These remain pillars of scientific research and in many ways they have remained the same. And yet, there are many changes in how these designs are implemented and research is conducted. Consider two features.

Experiments emphasize the random assignment of subjects to conditions. Random assignment distributes nuisance variables across groups, so these variables are unlikely to bias the results. This is exemplified in RCTs, the so-called gold standard for evaluating interventions. Yet, we want to evaluate interventions and programs in many situations in which random assignment is not possible. More often than not we cannot assign different interventions or experiences in such settings as schools, different clinics, and hospitals. Advances here include improved methods matching subjects in pre-formed groups even when the groups have not been randomly assigned. A family of matching procedures we discussed previously (propensity score matching) can match groups on many variables—over a hundred—for example. This is an enormous evolution in design and improves the ability to control for selection bias variables in new ways:

- *Does matching in this way solve all problems?*
- *Is matching better or worse than randomization?*

These are rarely the questions to ask in methodology.

You have learned in this text that any specific methodological practice has its own problems or trade-offs. Even randomization that I and other methodologists worship (and is spelled out it italics right above my butterfly tattoo) raises issues. For example, by definition randomly assigning subjects means that on occasion groups will be quite different on nuisance variables we wanted to distribute among groups. And, if the sample size is not very large (e.g., $N < 50$) even finding out if the groups are different will be difficult (e.g., low power). Also, in some contexts (e.g., treatment) subjects who are randomly assigned and end up in the "control" condition are more likely to drop out. There is nothing in my comments against randomization—would I have the tattoo if there were? Rather this is a commentary that no one practice is free from every concern. Improved matching in research allows stronger inferences from quasi-experiments where we begin with selection biases as a threat to validity. This threat can be made less plausible.

Running subjects is another change and advance in experiments. Psychological science has been heavily built on introductory college students (~67% of studies in the United States). We discussed increased evidence that inferences from college student samples have questionable external validity to other samples because college students, as subjects at least, are WEIRD (an acronym for Western, Educated, Industrialized, Rich, and from Democratic Cultures; Henrich, Heine, & Norenzayan, 2010a, b). Many core psychological processes often are influenced by cultural factors, and a broader range of samples is needed.

A **sea change** in running subjects is the use of online studies and reliance on technology. Now MTurk, Qualtrics, and Survey monkey have become familiar and prominent examples of how subjects can be obtained. Among the advantages is that the group is the public at large (although not necessarily representative of the population). Yet, this moves well beyond college students and will allow us to better identify whether and when subject characteristics moderate our findings and conclusions. Also, studies with online samples can obtain many subjects (increased power) in a very short period and the data obtained immediately enter into a database for analyses. It is now quite possible to "run subjects" without ever seeing the face of a real person or who comes into a lab and is met by a research assistant.

These are just two changes in the design and execution of experimental arrangements. I mention them to illustrate that experiments are evolving in how they are done. Let me illustrate other facets of methodology.

19.2.1: Assessment

We covered many different types of measures (e.g., objective, projective, psychological biological). Scientific progress depends so heavily on advances in assessment. So many topics and phenomena have been around for a while—virtually forever actually but we cannot get to them because we could not assess their presence or influence. For example, those little microbes in our gut (digestive system) and other places in our body out number our own cells by more than 10 to 1. Breakthroughs in assessment have allowed evaluation of their presence and roles (in learning, immune system, symptoms of autism and anxiety, with more to come).

At a much larger scale, we are "discovering" new galaxies and planets all of the time. Obviously, they are not new; they were for some time, but new to us. Novel assessments have emerged in telescope lenses, engineering, and math modeling to form those telescopes that have improved assessment. So we see all sorts of objects that we could not see before and at greater distances.

> Assessment provides knowledge of new realities and is a place where methodology leads to substantive advances.

For psychology, as I have mentioned, the improved and finer-grained version of neuroimaging with looking at networks in real time and examining the whole brain will be huge. Now core psychological processes (affect, cognition, learning, memory, perception, decision making) will be elaborated in entirely new ways as we track moment-to-moment changes in real time across individual neurons (e.g., Underwood, 2013). In terms of clinical psychology, these assessments too will be used to evaluate clinical states, symptoms, and disorders in new ways. We are troubled now by the fact we are trying to separate psychiatric disorders (e.g., clinical depression, anxiety) when so many disorders go together (comorbidity) and share common genetic features and risk factors (transdiagnosis). Finer-grained assessment is likely to change much of that. Advances in methodology, in the case assessment, will drive substantive advances and deepen our understanding of how similar paths (e.g., in early childhood) morph to diverse outcomes and how diverse paths morph into similar outcomes (e.g., depression, psychopathy, criminality).

Assessments in other ways will change research and our knowledge. Smartphones, tablets, smartwatches, and various wrist bands already are being used to assess individual states and experiences in real time in everyday life. Here technology has allowed for more in the field research. In addition, assessment can be connected to intervention. As emotional states, stress, or depression are evident from the assessment, self-regulation or coping skills can be automatically accessed on that same device. Sensing bodily states that correlate with psychological states is developing further as clothing is being designed to include measures. Here assessment advances will allow more research to capture processes throughout the day, in real time, and as participants live their daily lives. This will not replace experiments where individuals come to the lab, but opens further work opportunities outside of the lab.

19.2.2: Data Evaluation and Interpretation

Methods to evaluate data continue to evolve and expand, although these are slow to enter into most graduate programs in psychology. Novel and still unfamiliar statistics are needed for changing characteristics of much of the data that are being collected:

1. The scale of data is expanding. I previously mentioned the pooling of data sets (raw data from multiple studies) and big data. Increases in the scale of data require new statistical models and tests that will permit extracting relations.
2. More studies look at multiple variables as they relate to each other and change in real time. That is, increasingly we are less likely to look at the effect of variable x on y or even the effects of multiple variables as they relate to an outcome (e.g., as in regression analyses). Rather, we are more likely to look at multiple variables changing in dynamic ways and influencing each other.

> A better model to understand this is the weather and weather prediction in which events occurring in real time alter the predicted outcome in an ongoing way.

That is the moment-to-moment changes (or some real-time frame) have variables affecting each other and in the process changing the likelihood of possible outcomes. This is the way hurricane prediction can work. There is no plugging in of variables into an algorithm for a one-shot prediction but constantly changing variables that need to be "modeled" mathematically to see how these variables relate to each other and to the outcomes of interest. And so it is with most psychological phenomena. What a person does in life (e.g., career, marriage, volunteer to help others) may be influenced by several factors occurring at different points in time (e.g., childhood but other influences throughout that modify and moderate early influences including biological changes). Statistics are available that look at dynamic and interactive relations among variables (Little, 2013). As with other areas of specialization (e.g., neuroimaging), it is likely that more collaborative arrangements will be needed with statisticians and math modelers to incorporate advances on large data sets collected in real time.

A relatively common focus in clinical psychology research is to study mediators, i.e., those constructs that

might account for or explain the effect of an experimental manipulation or intervention. Mediation usually is studied by measuring a construct at before and after the manipulation and somewhere in the middle between those two and three fixed assessment occasions. The mid-assessment usually includes the proposed mediator. The stagnant and fixed nature of the "mid" assessment is limiting—it is not likely that all participants will change at the same point in time even if the mediator contributes to change. We need to observe ongoing changes and collect the data needed for that and then models (statistical, computational) to describe the processes of change. Ongoing assessment is needed but also statistical tests that can look at the data of individuals and groups.

19.2.3: Ethical Issues and Scientific Integrity

Ethical issues and scientific integrity, as we discussed, refer to a range of responsibilities and obligations researchers have in relation to participants, to science, and to the public at large.

Changes in how research is done require that guidelines and specific practices continue to evolve as well. Circumstances continue to emerge from advances in other facets of methodology (e.g., assessment) that were not specifically anticipated when the guidelines were originally formulated.

For example, pooling of data and big data already has raised consent and invasion of privacy issues. Big data often refers to the using technology and social media, tracking large amounts of information from people's daily life, and combining the data. Examples can include driving habits (from phone GPS) as people move around, credit card purchases, online Web browsing, e-mail, arrest records, and tickets for driving while under the influence of methodology. All of these and more are rich sources of information that can elaborate human behavior (e.g., in times of crisis, in response to local emergencies, social networking) and are making their way into psychological research. Often individuals are not identified in this research but groups, neighborhoods, and other such units can be. In such uses, consent is not solicited and obtained. There may be risk even so such as discrimination and negative stereotyping as groups, neighborhoods, and other such units are characterized. In these emerging applications, guidelines will need to develop to expand or consider when can data be shared, when is pooling and collecting data and its publication invasion of privacy?

What do you think?

The matter already is in the courts to consider what constitutes an acceptable practice and when privacy is invaded. In short, new ways of doing research and collecting data can require evolution of ethical guidelines to protect people who are participants in multiple projects of which they are completely unaware. Methods constantly evolve, and guidelines have to follow as new ethical risks are seen to come with them.

Informed consent raises new challenges and the need for new guidelines. For example, genetic information (kept in freezers for research) from individuals no longer living may reveal critical life and death health issues for their relatives who are living (e.g., Couzin-Frankel, 2014). More detailed analyses (from improved assessments) provide information that was not anticipated. The no-longer living did not provide consent for the unanticipated use. Should the information be shared with the living relatives beyond the use for which consent was granted? If yes or no, what will be the guiding principle?

Scientific integrity protections are undergoing change, and these are likely to continue. Very visible instances of scientific fraud in different scientific fields have converged to increase procedures and activities related to sharing of data, specification of details of studies in advance, promotion of replication studies, and publication of "negative" (no-difference) findings. Funding agencies and journal often require clarification of decision points in the collection, analysis, and reporting of data. The goal is to clarify and reduce questionable practices such as endlessly analyzing data in various ways to search for significance or to report only some of the many measures used in a study. In addition, increasingly data collected for a given study must be made available to others once a study is published and placed in the public domain. This will facilitate checking on the data, replication and extension of the data analyses to see if the conclusions remain, and replication of the entire study. As these examples illustrate, the broad responsibilities and obligations associated with scientific integrity have not changed. Yet the specific ways in which these responsibilities are met are changing very much.

19.2.4: Communication of Research Findings

Marked changes continue in communication of one's research findings particularly in journal publications. Changes in technology have facilitated many of these changes. Traditionally, publication of one's work appeared in a journal that was printed and circulated to other scientists and libraries. Increasingly, journals are not printed but rather are online. This has enormous implications and connects with other issues. For example, there are many reasons that "negative results" are not published and that there is a publication bias favoring studies that have found statistically significant differences and support for a hypothesis. Among them is that journals are limited in the number of printed pages they can produce in a given year.

Publication of an article in a print journal can be very expensive, and each page adds more expense. For most journals, financial profit comes from subscriptions of individuals and institutions (universities, libraries, industry). Even so, journals cannot add or keep adding pages to increase their length. If space for journal publication is very limited because of printed pages, naturally journals would give priority to studies that report finding something rather than finding nothing.

Yet, journals no longer are restricted to printed pages. Online publication has pushed aside and perhaps soon will completely dominate the printed journal. What that means is that there are no printed page limits and massive storage is available in the cloud or specific Web servers. Negative results could be much more readily accepted. In addition, data sets and detailed descriptions of experimental materials that in years past could not be published in a print journal also can be submitted with and connected to an online publication. Now materials that were not available in years past can be. Here is an instance in which changes in publication and communication are affecting other areas of methodology, including whether and how much information can be shared and increasing the possibility of replication.

Online materials including write-up of the original study and supplementary materials (data, descriptions of procedures) are very easily disseminated. Practices that are now antiquated (go to a library, find a journal, make a photocopy of the article) are all impediments to sharing of information because they are costly and time consuming. Now one can email a scientist a few continents away and within minutes have pdf files of the article and all the materials. Communication of findings among scientists has and is changing greatly.

19.2.5: General Comments

In these comments, I have only sampled facets within the areas of methodology that are evolving.

The overall point: *Methodology is constantly changing and evolving.*

This helps convey an earlier point I have made that methodology is not merely a series of practices that one can learn, apply, and be finished with that. It is important to have a deeper understanding of what methodology is designed to accomplish (e.g., drawing valid inferences, making competing interpretations implausible) because these broader features help one adapt to the new circumstances of experimentation.

No text could keep up with the advances in each of the areas I have delineated. Similarly, no individual researcher could keep up either. Yet, that is not necessarily the task.

> Collaboration of colleagues from different areas of psychology and from different disciplines is a key strategy to draw on novel theory and methods.

Disciplines that might seem far afield often offer remarkable contributions because of the diverse perspectives and methods they use. For example, in clinical work, we think of "treatment" as based on psychological interventions (e.g., various forms of therapy) and medication for psychological dysfunction (e.g., addictive behavior, depression). Yet a vast array of interventions drawing on practices (e.g., leisure activities, spiritual experiences, meditation) and disciplines (e.g., economics, public health, policy) not routinely used or taught in psychology influence clinical dysfunction (e.g., Kazdin & Rabbitt, 2013; Walsh, 2011). Collaboration with others interested in a particular target focus (e.g., clinical dysfunction, health, drug use among teens) allows one to expand methods of study and hence the range of insights likely to be revealed.

19.3: Importance of Methodological Diversity

19.3 Evaluate the importance of multiple scientific research methodologies

For several topics in the text we have discussed different options and approaches. For examples at the broadest levels, three different research traditions were discussed including:

- Quantitative
- Qualitative
- Single-case research

Also, at a broad level two approaches to data evaluation were discussed:

- Statistical methods
- Nonstatistical methods

And of course within a given research tradition (e.g., quantitative research), there are many design options (true-experiments, quasi experiments, observational studies) and approaches to hypotheses testing (e.g., null hypotheses, Bayesian) were discussed. The range of options for research is remarkable.

Diversity of methodological approaches is critically important. To begin with, the relations among variables can vary as a function for how they are studied. We know this in an obvious way. Two researchers studying aggression or domestic violence might yield different findings based on how they operationalize these constructs. But more than that, different methods of studying a phenomenon can yield different findings.

For example, in clinical psychology, we want to know the life-time prevalence rates of various psychiatric disorders (e.g., depression, obsessive compulsive disorder).

We have learned the rates are grossly underestimated from retrospective studies when compared to longitudinal studies that assess clinical dysfunction over time (see Takayanagi et al., 2014). Interestingly, in this same study, the rates of physical disorders were much consistent across retrospective and longitudinal designs.

More generally cross-sectional and longitudinal designs of a given phenomenon (e.g., cognitive decline with age, effects of not monitoring the whereabouts of teens) show similar effects, but the magnitude of the effects sometimes varies markedly as a function of which design is used (e.g., Lac & Crano, 2009; Salthouse, 2010). Also, conclusions reached about brain activation associated with practice can vary on how the data are analyzed and whether the emphasis is on group analyses versus individually based analyses (e.g., Ganis, Thompson, & Kosslyn, 2005).

The overall point rather than specifics of these examples is the most critical. How something is studied and evaluated can influence the results and conclusions. This does not mean one should distrust, dismiss, and discount a finding evaluated in one way. Rather, our understanding is enhanced by evaluating the phenomena of interest in different ways and seeing whether conclusions will vary as a result.

> When our conclusions do vary based on how they are studied, that more often warrants celebration rather than despair. Why are there differences?

That answer deepens our understanding of the phenomenon or of methodology.

Also, different methodologies give different levels of analysis. Group research focuses on means and standard deviations. From quantitative group research, we can identify critical influences and relations among variables (e.g., risk factors for posttraumatic stress disorder). At a different level and from qualitative research, we can get an idea of precisely how individuals experience trauma. The in-depth analysis of qualitative research often without structured questionnaires can identify commonalities and differences among individuals un-going trauma. From in-depth descriptions, themes, and the structure as experienced can be identified. Arguably, really informed questionnaires and measures can be developed from a more personal and thorough understanding of what we wish to assess. Single-case designs drill down and can evaluate treatment rigorously to alleviate the impact of trauma. With single-case data, we are not looking for group means among all individuals who receive the intervention, but the impact on individuals.

As individual researchers, most of us specialize in a topic or two. As part of that specialization, we usually conduct a narrow range of designs and use a narrow range of measures and data-analytic techniques. This is natural and required for specialization, and in this context narrowness is not necessarily negative. Yet, for the science as a whole across many different researchers the diversity of approaches is extremely important.[2]

Think of methods of research (different designs, assessments, and so on) as a lens or rather a set of lenses through which we study, view, and understand natural phenomena. The results we obtain depend very heavily on the lens we use. Let me convey the point with "real" lenses. For decades, the National Aeronautics and Space Administration (NASA) in the United States, in collaboration with other countries, has had a Great Observatories Program, which includes different telescopes out in space (NASA, 2009). The different telescopes look at the full electromagnetic spectrum or wave lengths, including the spectrum that is visible to us but also gamma rays, X-rays, and infrared. (The most familiar is the Hubble Space Telescope, but the three others in the program have included the Compton Gamma Ray Observatory, Chandra X-Ray Observatory, and Spitzer Space Telescope.) This program is now decades old, and evolving as upgrades and replacement telescopes expand the program and its capabilities. For the present discussion, the critical point is that observing the universe in different ways has yielded quite different findings. The telescopes when pointed to the same object show different pictures and provide different information. Needless to say, no one view from one telescope is better or more accurate; each reveals the reality to which it is sensitive. Any one telescope would be limiting.

And so it is with methodological traditions and practices discussed in this text. We want diversity of approaches precisely because what one sees depends on how one is looking and through what lens. We want all of the methodological approaches brought to bear to understand core psychological processes (e.g., cognition, perception, and sensation) and also critical topics within clinical psychology (e.g., risk, resilience, emergence, and etiologies of psychiatric disorders).

19.4: Abbreviated Guidelines for a Well-(and Quickly) Designed Study

19.4 Review the guidelines to effective statistical methodologies

The text has discussed methodology as a way of thinking. In addition, the components of methodology (research design, assessment, data evaluation, ethical issues, and publication and communication of findings) were reviewed in some detail. While all of this material is important, many of us long for something truly concrete that will actually help us. By analogy, it is nice to know how my smartphone or tablet works, but at some point I just want to know how to watch a movie.

Understandably, the reader who has arrived this point might identify lingering practical questions that have not been answered or answered very well, namely:

How do I design my study right now?

Where do I begin?

What are the issues to which I have to attend?

The answers are in the previous chapters, but perhaps they have to be mined to find a few gems hidden in tons of rock. Methodology can be simplified to aid the reader with the practical questions.

I have mentioned that designing and completing a study and communicating the results are very much like a story. There are beginning, middle, and ending parts. They convey a theme and at once bring the story line to a conclusion but also to some yet-to-be-resolved future outcome. The idea of a story can help the researcher plan, execute, and write up a study. In that spirit, this Appendix is designed as an aid to help the interested reader in concrete ways to move rapidly from pages of methodology to a study.

Table A.1 presents a story outline for you to use. Here is the recommended use. Take the story outline (the table) in a room where there is privacy and you can talk out loud without being threatened with hospitalization or heavy medication. (Although I have no data, personal experience suggests that using this in public places has rather odd interpersonal consequences. I tried this in a well-known coffee shop chain and then in a fast-food restaurant with a Scottish name and, trust me, few people in either setting were interested in methodology, to say the least. So the usual recommendation "do not try this at home" does not apply here—home may be the only place to try this.) As you talk, merely complete the multiple-choice questions (circle the choices in parentheses) or fill-in the options (as indicated by a blank space underlined in the text) to complete the story. When you complete the story, you ought to have a well-designed study! See how simple methodology can be? Makes one wonder why I did not just begin the text with this story outline and leave the rest as an appendix.

Guidelines for Developing a Well-Designed Study

Directions: The story is designed to be read aloud by the investigator in a private setting. Read each sentence slowly and complete the questions. There are multiple-choice questions, indicated by parentheses, where you are required to circle or underline one of the options and fill-in questions, indicated by a blank underline, where you are required to write in what you will do. Years of use of these guidelines have shown that the quality of the final study is deleteriously affected if the answer format is violated (i.e., if the investigator writes in answers for the multiple-choice parts and circles the blank underlines). Please begin reading slowly and mellifluously now. (For those of you with theater experience, please commit this to memory and as you recite this your motivation is "inquisitive joy." You love methodology and your parents read about assessment, design, and statistics instead of the usual child fare. Bring that out in your performance.) Ok, you are ready.

"Well I am finally going to begin a study. This going to be an important study because other studies that even come close to this have not [fill in]. Probably this study, if I ever get it done, will contribute to (the knowledge base, science, humanity, my career [circle one]) in at least two ways 1) and 2), and I know I am being modest. The field is really fortunate to have me to do this study. I am a pretty amazing person. My (mom, dad, significant other [circle one]) certainly got that part right.

I know I have to prepare a proposal with the details of the study for (the Institutional Review Board, my advisor, yet one more deadline for the graduate school or grant agency [circle one]). This will be easy. As background to my study I (thoroughly reviewed, briefly glanced, decided to ignore [circle one]) prior theory and research. But I am ready to design the study.

Basically, I have these (2, 3, 4) (predictions, hypotheses) and they are:

1) _____

2) _____ (add 3, 4 as needed)

I even imagine how I shall analyze the data to test these. Probably to test these I shall use _____ (list some statistical tests or analysis). Of course this is just tentative but I am pretty cool to even think of the data analyses at this point.

The subjects for this study are going to be _____. I chose this sample, not just because they are convenient or because my advisor or colleague has them around—everyone does that. I admit, actually was tempted. But hey, no, not me, I am using these subjects because _____. I am going to use lots of subjects and in fact my sample size (God willing) will be _____. Of course I did not pull this number out of the air. I looked at a Web site for the search word "power." Yeah, first I got a bunch of electrical tools like drills and saws but I redid this (I am a pretty thorough person) typing in "statistical power." In a few minutes I was able to estimate the sample size and looks like I will need _____ number of subjects. I estimate I will need a sample of this size to have a chance to find differences if they exist. To be honest, I hope this sample helps me find differences even if they don't exist. A big sample can't hurt either of these goals.

For the subjects to be included they have to meet these (2, 3, 4) criteria: 1) _____, 2) _____, and 3) _____. I Will exclude them if the meet these (2, 3, 4) criteria 1) _____, 2) _____, and 3) or have an "attitude problem." To measure these criteria, I will look at them, ask them a few questions, or give them these measures _____.

Speaking about measures, I am going to use a lot. Actually, I care about these constructs: _____; _____ (add _____ as needed). For the first construct these

measures will be used _____; for the second construct these measures will be used _____ (etc.). I am also throwing in this(ese) measures because they: (are interesting, are used a lot in this area, may explain why the results come out the way they will, are being a pushed by my advisor, colleague, mother [circle one]).

The main experimental manipulation treatment I will be using is _____. I will be using (guidelines, manual, book on treatment [circle one]) that I (got from a researcher who developed this manipulation or treatment, a credible imitator; or I invented myself [circle one]). This is a reasonable version of the manipulation. If this is a treatment study, the treatment will be given to participants for a period of _____ hour(s), for _____ weeks at about _____ sessions per week. I am going to train the therapists, establish criteria to decide when they are trained, and then to monitor the delivery of treatment during the study. How am I going to do the monitoring? Well, I plan to: _____ and _____. I will get some measures of treatment integrity to see that my efforts are not in vein and to see if I or anyone else ought to believe the results. If this is not a treatment study, I will still check on the manipulation to see if the subjects grasped what my manipulation was designed to accomplish.

Oh. I almost forgot. In this will be a (between-group experiment, observational study) and I plan on having (fill in #) _____ (groups or subjects). As applicable if a group study—The groups include _____. There will be (0, 1, 2) control groups and these include _____ _____. Many people just throw in a control group without being clear as to why. Not me. My control groups are designed to control for threats to _____ validity and of course are essential to test the hypotheses.

Here is what happens to a subject when he or she comes to this project. First, we give a big (interview, welcoming speech, assessment battery, hug [circle one]) then of course seek informed consent. Yes, the university review committee that looks at ethical issues has already approved my study, and I have the consent forms finally resolved to pass muster. My God, getting the wording right and obtaining final approval for the consent forms were (bizarre, no picnic, a breeze because I copied my advisor's forms [circle one]). Then the subject will (complete, come back for [circle one]) the assessments. I will then assign subjects (randomly, as the heart may prompt [circle one]) to conditions. The experimental and control conditions now are implemented and then followed by posttreatment assessment (right after the last session, on the same day, within one week).

When the data are in, I will look at basic characteristics of the sample, look at these by groups, and see if the means and standard deviations look reasonable. The data were all checked and entered twice so I am pretty sure there are no big errors. Even so, it is good just to look at what the sample and individual groups are like.

My advisor is a little rigid about dumping so-called outliers, so I will keep everyone in the study unless he or she died or something like that. Just as a learning experience, I may peek at the data to see how the results change when wildly odd subjects (3 standard deviations from the mean) influence the data.

As is my style, I shall probably analyze the data with every statistic I have ever learned and no doubt click my mouse on a few that I have no idea about. Hey—how can one learn without trying new things? But, I shall provide very very specific tests of my hypotheses with focused statistical tests and present these so that the reader can see the hypothesis, the tests, and my conclusions. A huge issue now relates to questionable research practices like running a zillion statistical tests or reporting just a few of the many measures. I will be open about this and reveal all (in a footnote, supplementary document, at confessions [circle one]).

As for the write-up, clarity is not my forte and people have been on me for that. If the results show anything else interesting including possible confounds, I shall present that too but probably sequester that (whatever that means) from the section that gives the main findings. The reader will be confused if I am too. Maybe a separate heading called (Supplementary Analyses, Exploratory Analyses, or Huge Fishing Expedition [circle one]) will be the place to note these other analyses and findings.

There will be so much richness and depth to my study, and my work in general, that I probably ought to begin my discussion of the results with a brief overview or statement of the main findings. Probably people reading the results of my study such as (blood relatives, Nobel committee members, authors looking for great examples of research for their methodology textbooks [circle one]) will like a clear summary opening paragraph. After that, I shall try to (describe, explain [circle one]) a key finding or two in more detail. As soon as I can, I shall make comments about (how this relates to, builds on, trumps, completely shames [circle one]) other work that has been done on the topic. If there is any (theory, other areas of research outside my topic [circle one]) to which I can relate the study, I shall (toss, squeeze, force [circle one]) that in as well.

I also will make a few comments about the limitations of my study. Given how I have designed this study and my own personal skills, even if I say so myself (and I often do), this could be a very brief limitations section. But no study is perfect, and real limitations are always present. In writing this section, I shall try not to (get too defensive, righteous about how I chose to design the study, attack the reader's possible skepticism [circle one]). In the remote chance that there are serious limitations of the study, more likely than not I shall (blame my advisor, remind readers of my difficult childhood, use a small font for that part of the Discussion [circle one]).

Finally, if space allows, I shall talk about the next study that ought to be done to build on my work. This future work ought to be an important study and not merely a test of generality to a different sample or setting unless there is some special reason to suspect generality would be an issue. Something really (meaty, inspiring, conceptually interesting [circle one]) in this paragraph will suggest a new "story" that needs to be told. Who knows, maybe I'll even (do that study, rest on my laurels [circle one]) after completing this study.

Summary and Conclusions: Methodology

The topics of clinical, counseling, educational, and school psychology and other areas that combine basic science with application are critically important. They bring to bear diverse types of studies, including laboratory experiments with human and nonhuman animals, the studies in institutional settings (e.g., schools, prisons), cultures, and clinical samples. Consider some of the topics within clinical psychology, such as:

- Etiologies
- Course and recovery of mental illness, addictions, self-injury, and suicidal behavior
- Clinical dysfunction over the course of development
- Role of stress on adjustment, parenting, relationships, and clinical dysfunction
- Relation of physical and mental health
- Interpersonal violence in its many forms (e.g., child abuse, domestic violence, date rape), including victimization and perpetration
- Treatment, prevention, and alleviation of the many forms of clinical dysfunction
- Subjective well-being, happiness, and social relations (e.g., family, partners, spouses, friends)

Of course, this is barely the tip of the iceberg. The topics have no one feature in common. Well perhaps they do. Each is a complex, multifaceted topic with scores of influences and consequences and potentially confounding variables.

The methods we have discussed in this text are the path to understanding and separating artifact, bias, and distractors from the influences and how they operate. Elaborating complex phenomena (e.g., domestic violence, schizophrenia, happy marriages) is like finding a needle in a haystack.

> Methodology is sort of like a magnet that can attract needles, separate them from the hay, and permit us to describe and explain what we need to know.

Methodology is not an end in itself. The goal of psychological science is to understand the many topics within our domain, as sampled above, so we know how, why, and what to do to improve public life and the environment in which life is supported. Sound methodology works in conjunction with developing theory and hypotheses. Methodology is a critical partner in obtaining the knowledge we seek. The goal of the text was to provide initial stages in developing mastery and fostering a broader appreciation of the contribution of methodology to the substantive topics of clinical and counseling psychology, to related social sciences, but of course science more generally.

Critical Thinking Questions

1. What is methodological diversity, and why is it important?
2. What are two examples from any area of science where advances in assessment have helped make an advance in our understanding of something about the world?
3. An easy final question: What are the five areas of methodology covered in this text?

Chapter 19 Quiz: Methodology: Constantly Evolving along with Advances in Science

Glossary

ABAB Design A single-case experimental design in which the performance of a subject or group of subjects is evaluated over time across baseline (A) and intervention (B) conditions. A relation is demonstrated between the intervention and performance if performance changes in each phase in which the intervention is presented and reverts to baseline or near baseline levels when it is withdrawn. Also called Reversal design.

Accelerated, Multi-Cohort Longitudinal Design A prospective, longitudinal study in which multiple groups (two or more cohorts) are studied. Each group covers only a portion of the total time frame of interest. The groups overlap in ways that permit the investigator to discuss the entire development period from the youngest through the oldest cohort, although no individual cohort covered that entire span.

Alpha (α) The probability of rejecting a hypothesis (the null hypothesis) when that hypothesis is true. This is also referred to as a Type I error.

Alternate-Form Reliability The correlation between different forms of the same measure when the items of the two forms are considered to represent the same population of items.

Anecdotal Case Study Intensive study of the individual in which there is no systematic or objective assessment. The "data" (information) are based on narrative and informal reports of the client and therapist without checks on their reliability or validity. See **Case Study** and **Single-Case Experimental Designs**.

Anonymity Ensuring that the identity of the subjects and their individual performance is not revealed. Participants who agree to provide information must be assured that their responses are anonymous in order to protect their privacy.

Archival Records Institutional, cultural, or other records that may be used as unobtrusive measures of performance.

Artifact An extraneous influence in an experiment that may threaten validity, usually construct validity.

Assent Children and adolescents who participate in research must agree (provide assent) to participate if they are old enough to understand the proposed research, activities expected of them, risks, and benefits. For the assent criterion to be met, the child must affirmatively agree to be involved in the research project. Assent is in addition to rather than instead of informed consent provided by a parent or guardian.

Attention-Placebo Control Group A group in treatment research that is exposed to common factors associated with treatment, such as attending treatment sessions, having contact with a therapist, hearing a logical rationale that describes the genesis of one's problem, and so on. The group may receive a "fake" treatment, i.e., an intervention that provides the common factors without providing an intervention that on theoretical or empirical grounds would be expected to be therapeutic.

Attrition Loss of subjects in an experiment. This usually means dropping out before the study is completed. The loss of subjects can threaten all facets of experimental validity.

Baseline Assessment Initial observations used in single-case designs that are obtained for multiple occasions (e.g., several days) prior to the intervention.

Baseline Phase The initial phase of most single-case experimental designs in which performance is observed on some measure for several occasions (e.g., days) prior to implementing the experimental condition or intervention.

Bayesian Data Analyses An approach to statistical analyses that take into account prior knowledge or information to estimate probability of a particular outcome. The analyses provide an alternative to null hypothesis statistical testing. Rather than trying to reject the null hypothesis, Bayesian analyses pit the null hypothesis against the alternative hypothesis to evaluate which is more likely given the data. There are multiple ways of conducting Bayesian analyses and models of estimating and evaluating probabilities on which the analyses depend.

Behavioral Measures Assessment that focuses on overt performance in laboratory or everyday settings. The performance attempts to sample directly the behavior of interest.

Bench to Bedside A term used in translational research to refer to moving a finding from basic, laboratory (bench) to clinical application with patients (beside). This can be thought of as moving from basic research to application.

Beta (β) The probability of accepting a hypothesis (the null hypothesis) when it is false. This is also referred to as a Type II error.

Big Data Refers the harnessing of massive amounts of information that is available and utilizing that in novel ways. Big data is not just more data. The amount of data and the integration from several sources (e.g., social media, health records, real-time processing of neurons, networks of the brain) require novel technical challenges and evaluative strategies.

Birth-Cohort Study A prospective longitudinal study that begins with a group of subjects who enter at birth. Usually a specific time frame (e.g., 6- or 12-month period) and geographical locale (country, island, state or province, district, hospital) are identified. Children born in the specific time period and geographical setting serve as participants and then are followed for an extended period through childhood and adulthood.

Blind A term used to denote a procedure in which the experimenter and others associated with the investigation (e.g., staff, assessors) are kept naive with respect to the hypotheses and experimental conditions. Because of the confusion of the term with loss of vision and the pejorative reference to that condition, terms other than "blind" (e.g., masked, experimentally naive) are preferred. However, with its long history, "blind" continues to be used frequently in methodology.

Bonferroni Correction Refers to a way of controlling for the probability of a Type 2 error when several multiple comparisons are completed. The alpha (p level) is adjusted for the individual comparisons to control for the overall error rate.

Buffer Items Items or content of a scale or measure that are intended to disguise or dilute the focus of interest evident in the measure. For example, items related to hobbies or physical health in a self-report scale on psychopathology might be added to serve as buffer or filler items.

Carryover Effect In multiple-treatment designs, the impact of one treatment may linger or have impact on a subsequent treatment. This is equivalent to multiple-treatment interference.

Case Study An intensive evaluation and report of an individual subject. In psychology, this usually means one person, but a case study can focus on larger units (e.g., one city, one business) See **Anecdotal Case Study** and **Single-Case Experimental Designs**.

Case-Control Design An observational research design in which the characteristic of interest is studied by selecting individuals to form groups. The groups vary on that characteristic (e.g., depressed vs. not depressed). Once the groups are formed, other current or past characteristics (e.g., family relations, personality) are studied. Minimally two groups are included, namely, those who show the characteristic of interest (cases) and those who do not (controls).

Cause or Causal Relation In science, a causal relation is drawn between two or more variables when several conditions are met. These include a strong association between the variables of interest, consistency or replication of that association, specificity showing a clear connection between one variable or set of variables and outcome, a clear time line where one variable becomes before the other, and experiment or showing intervening in one variable alters the other, as well as other criteria. The most familiar and relied on criterion for inferring cause in experimental research is showing that a phenomenon can be altered by manipulating the variable considered to be a cause.

Ceiling Effect This refers to an upper limit in the range of scores of a measure. The limit may preclude the ability to show differences among alternative groups or conditions. The effect may be especially likely in multiple-treatment designs. As treatments are added or as the client has changed from a prior treatment, there may be little room (on the measure) to reflect incremental benefits of treatment. Ceiling or floor effect is used as a term depending on whether the upper or lower limit of the scale provides the restriction.

Certificate of Confidentiality A further layer of protecting privacy of participants issued by the National Institutes of Health. The certificate allows the investigator and others who have access to research records to refuse to disclose identifying information on research participants in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level. This certificate often is used in clinical trials where clients participate and disclose sensitive personal information.

Changing-Criterion Design A single-case experimental design that demonstrates the effect of an intervention by showing that performance changes in increments to match a performance criterion. That performance criterion changes at different points throughout the intervention phase to show that performance responds to the change in the criterion.

Checking on the Manipulation Refers to assessing the independent variable and its effects on the subjects in ways that are separate from the effects of the manipulation on the dependent variables. The assessment is to see whether the manipulation "took" or altered what was intended (e.g., mood, attitude).

Clinical Significance The extent to which the effect of an intervention makes an "important" difference to the clients or has practical or applied value. This is most commonly measured by normative comparisons, reliability of change, and no longer meeting criteria for a psychiatric diagnosis that may have been required to be included in the study.

Cognitive Heuristics Processes out of our awareness that help us organize and integrate information. They serve as mental short cuts and help us categorize, make decisions, and solve problems. They introduce bias when we try to draw accurate conclusions based only on our own thoughts, impressions, and experience.

Cohort A group of subjects followed over time who share a particular characteristic. The usual use of this is in age cohort. Groups of different generations would represent different age cohorts.

Cohort Design An observational research design in which the investigator studies an intact group or groups over time, i.e., prospectively. The design is also referred to as a prospective, longitudinal study.

Completer Analysis A way of analyzing the results of a study that includes only those subjects who have completed treatment in a clinical trial. Subjects who have not completed the measures (e.g., who dropped out of treatment before posttreatment or follow-up assessment) are omitted from the data analysis. Contrast with Intent-to-Treat Analysis.

Conceptual Replication A study that tries to reproduce the primary relationship, concept, or principle of the original study but in a very different context. This would be a test that abstracts the principle or guiding concept of a study and applies this in a way that typically uses quite different methods (procedures, subjects) of the original study from which it draws.

Concurrent Validity The correlation of a measure with performance on another measure or criterion at the same point in time.

Confederate A person who works as an accomplice in the investigation, although he or she appears to be another subject or part of the natural arrangement of the setting (e.g., someone in a waiting room).

Confidence Interval A range of values (upper and lower) that reflect the likelihood that the difference in the population falls within a particular range. The range is based on estimates from the sample data in the same way that is used for evaluating statistical significance testing. Common values used for confidence intervals are 95% or 99%, which parallel statistical criteria for alpha of .05 and .01.

Confidentiality Refers that information in a study will not be disclosed to a third party without the awareness and consent of the participant.

Confirmability A criterion invoked to evaluate data in qualitative research and refers to the extent which an independent reviewer could conduct a formal audit and re-evaluation of the procedures and generate the same findings.

Conflict of Interest In relation to research, any situation in which the investigator has more than one role, incentive, or relationship to the procedures or goals of the project. A conflict would be evident, for example, if the investigator were interested in evaluating the impact of the intervention (role of the scientist) and at the same time were interested in the success of the intervention in light of possible financial gain (role of the entrepreneur). When an investigator holds stock in a business that may gain from the findings or when the findings might be used now or in the future for some financial gain, conflict is evident. Conflict of interest or the appearance of a conflict of interest is central issues in ethical issues of conducting research.

Confound A factor, other variable, or influence that covaries with the experimental condition or intervention.

Construct Validity In the context of experimental design, this refers to a type of experimental validity that pertains to the interpretation or basis of the effect that was demonstrated in an experiment. In the context of psychological assessment, the term refers to the extent to which a measure has been shown to assess the construct (e.g., intelligence) of interest. This latter use of construct validity requires multiple studies whose results are in keeping with what would be expected of the construct.

Content Validity Evidence that the content of the items of a measure reflects the construct or domain of interest. The relation of the items to the concept underlying the measure. This is evaluated by seeking the opinions of experts regarding the content as well as statistical procedures (e.g., factor analysis) that examine how many items go together and whether the items appear to be well represented by the construct intended.

Convergent Validity The correlation between measures that are expected to be related. The extent to which two measures assess the similar or related constructs. The validity of a given measure is suggested if the measure correlates with other measures with which it is expected to correlate. Contrast with Discriminant Validity.

Counterbalanced A method of arranging conditions or tasks to the subjects so that a given condition or task is not confounded by the order in which it appears. If two experimental conditions (A,B) are

to be evaluated, one group would receive A and then B; the other group would receive B and then A, so the conditions can be said to be counterbalanced.

Credibility A criterion invoked to evaluate data in qualitative research and reflects whether the methods and subjects are appropriate to the goals and are likely to represent the sample of interest.

Criterion Validity Correlation of a measure with some other criterion. This can encompass concurrent or predictive validity. In addition, the notion is occasionally used in relation to a specific and often dichotomous criterion when performance on the measure is evaluated in relation to selected groups (e.g., depressed vs. nondepressed patients).

Crossover Design A design in which two interventions are presented to each subject at different points in time. Halfway through the investigation, each subject is shifted to the other intervention or condition. The intervention is evaluated by comparing subject performance under the separate conditions.

Cross-Sectional Design The most commonly used version of a case-control design in clinical psychology in which subjects (cases and controls) are selected and assessed in relation to current characteristics. This is to be distinguished from studies that are designed to evaluate events or experiences that happened in the past (retrospective studies) or the future (prospective studies).

Data-Evaluation Validity The extent to which a relation between independent and dependent variables can be shown based on some facet of the data evaluation such as excessive variability and weak statistical power. A term for this has been statistical conclusion validity but was replaced in this text because not all data problems that threaten validity of a study involve statistical tests or issues.

Debriefing Providing a description of the experiment and its purposes to the subject after the investigation when deception was used or information was withheld about the investigation. The purpose is to counteract or minimize any lingering negative effects of the experimental manipulation.

Deception Presentation of misleading information or not disclosing fully procedures and details of the investigation.

Demand Characteristics Cues of the situation associated with the experimental manipulation or intervention that may seem incidental but may contribute to, or even account for, the results.

Dependability A criterion invoked to evaluate data in qualitative research and pertains to the reliability of the conclusions and data evaluation leading to these conclusions.

Dependent Variable The measure(s) designed to reflect the impact of the independent variable, experimental manipulation, or intervention. Contrast with Independent Variable.

Diffusion or Imitation of Treatment The inadvertent administration of treatment to a control group, which diffuses or obscures the impact of the intervention. More generally, any unintended procedure that may reduce the extent to which experimental and control conditions are distinct. This might also occur if someone in the intervention group does not receive the treatment or receives the condition provided to control subjects.

Direct replication A study designed to repeat a prior experiment using methods that are maximally similar to those used in the original study.

Directional Test In hypothesis testing, some predictions are clearly directional where one group is predicted to be better (or worse) on the dependent variable of interest and the investigator has no interest in testing whether groups different in either direction (better or worse). A directional statistical test (e.g., *t* test) uses one rather than both tails of the distribution from which an inference is drawn as to whether statistical significance has been achieved. A directional or one-tailed test requires a lower level (smaller *t* test) to reject the null hypothesis. Arguably, most significance testing might be one-tailed but this is infrequently done.

Discriminant Validity The correlation between measures that are expected not to relate to each other. The validity of a given measure is suggested if the measure shows little or no correlation with measures with which is expected not to correlate because the measures assess dissimilar or unrelated constructs. Contrast with Convergent Validity.

Double-Blind Study A procedure often used in medication trials in which the patients (subjects) and those who administer the drugs (physicians or nurses) are not informed of whether they are receiving the medication or a placebo. The goal is to reduce the likelihood that expectancies or knowledge of the condition, rather than the effects of medication, could influence or account for the results.

Dual Use Refers to research that might provide knowledge, products, or technology that could be directly misapplied by others and could pose a threat to public health and safety, agricultural crops and plants, animals and the environment, or national security. The same findings or procedures could be used to enhance (e.g., improved health) or undermine (e.g., bioterrorism) public life.

Effect Size A measure of the strength or magnitude of an experimental effect. Also, a way of expressing the difference between conditions (e.g., treatment vs. control) in terms of a common metric across measures and studies. The method is based on computing the difference between the means of interest on a particular measure and dividing this by the standard deviation (e.g., pooled standard deviation of the conditions).

Effectiveness The impact of treatment in the context of clinical settings and clinical work, rather than well-controlled conditions of the laboratory. In effectiveness studies, treatment is evaluated in clinical settings, with clients as usually referred and therapists who usually provide services, and without many of the rigorous controls of research. Effectiveness and efficacy studies can be considered to reflect a continuum of experimental control over several dimensions that may affect external validity of the results. Contrast with Efficacy.

Efficacy The impact of treatment in the context of a well-controlled study conducted under conditions that depart from exigencies of clinical settings. Usually in efficacy studies, there is careful control over the selection of cases, therapists, and administration and monitoring of treatment. Contrast with Effectiveness.

Experimenter The person who conducts the experiment, runs subjects, or administers the conditions of research. See also **Investigator**.

Experimenter Expectancies Hypotheses, beliefs, and views on the part of the experimenter that may influence how the subjects perform. Expectancy effects are a threat to construct validity if they provide a plausible rival interpretation of the effects otherwise attributed to the intervention.

Experiment-Wise Error Rate The probability of a Type I error for all of the comparisons in the experiment, given the number of tests. Contrast with Per Comparison Error Rate.

External Validity The extent to which the results can be generalized or extended to persons, settings, times, measures, and characteristics other than those in this particular experimental arrangement.

Face Validity The extent to which a measure appears to assess the construct of interest. This is not regarded as a formal type of validation or part of the psychometric development or evaluation of a measure. The fact that a measure may appear to measure a construct of interest does not mean that it does or does very well.

Factorial Designs Group designs in which two or more variables are studied concurrently. For each variable, two or more levels are studied. The designs include the combinations of the variables (e.g., 2 *x* 2 design that would encompass four groups) so that main effects of the separate variables as well as their combined effect (interactions) can be evaluated.

Falsifiability The view underlying null hypothesis testing in which the results of a study can be used to falsify the null hypothesis (no

differences) rather than to prove an alternative hypothesis. A statistically significant finding is used to suggest that the null hypothesis can be rejected.

File-Drawer Problem The possibility that the published studies represent a biased sample of all studies that have been completed for a given hypothesis. The published studies may reflect those that obtained statistical significance (i.e., the 5% at the $p < .05$ level). There may be many more studies (the other 95% somewhere in a file drawer) that did not attain significance and were never published.

Fraud Explicit efforts to deceive and misrepresent. Altering or faking data and providing misinformation to deceive are the primary examples.

Ghost Authorship Refers to someone writing up a study in whole or in part but is not named as an author or noted in the acknowledgment section. This has emerged as a special problem in pharmaceutical research where contributors who are well known are listed as authors but in fact someone else not credited at all completely wrote the article.

Gift Authorship See **Honorary Authorship**.

Global Ratings A type of measure that quantifies impressions of somewhat general characteristics. Such measures are referred to as "global" because they reflect overall impressions or summary statements of the construct of interest.

Grounded Theory A term used in qualitative research to reflect the development of theory from careful and intensive observation and analysis of the phenomenon of interest. The abstractions, themes, and categories that emerge from intensive observation are grounded in and close to the data of the participants' experiences.

Health Insurance Portability and Accountability Act (HIPAA) A federal Act in the United States that is designed to ensure the privacy of client health information. Privacy includes the individual's right to control access to and disclosure of health information provided by the patient. Health information is defined broadly and includes matters related physical and mental health, psychological problems, and special services of other types (e.g., special education programming).

Healthy Controls A term used to refer to subjects who are from the community recruited because they do not meet the criteria for the dysfunction or disorder that is the main focus of the study. Thus, the study compares individuals with some characteristic (e.g., depression, bipolar disorder) to those who have no dysfunction, i.e., are "healthy."

Hello-Good-Bye Effect In the context of psychotherapy research, changes in self-report responses before and after therapy may reflect exaggeration (hello) at the beginning of therapy and underplaying of the problems (good-bye) when therapy is completed, rather than any true improvements in the referral of symptoms. This cannot usually be separated from other influences (e.g., testing, statistical regression).

History A threat to internal validity that consists of any event occurring in the experiment (other than the independent variable) or outside of the experiment that could account for the results.

Honorary Authorship Refers to individuals who are added to the list of authors on a manuscript but who have not contributed to the conception and design of the study, the collection, analysis, interpretation of the data, and drafting of the article. Also called gift authorship.

Imputing Data A way to handle missing data points by estimating what the data points would be based on equations that draw on other from subjects without missing data. There are multiple models (equations) that can be used to estimate the missing data points.

Incremental Validity Refers to whether a new measure or measure of a new construct adds to an existing measure or set of measures in predicting some outcome. That outcome might be in the present or future. Incremental validity is evident if the new measure adds significantly (statistically) to another set of predictors or measures.

Independent Variable The construct, experimental, manipulation, intervention, or factor that whose impact will be evaluated in the investigation. Contrast with Dependent Variable.

Informants Persons in contact with the client such a spouse, peers, roommates, teachers, employers, friends colleagues, and others who might be contacted to complete assessment or to provide information.

Informed Consent Agreeing to participate in research with full knowledge about the nature of treatment, the risks, benefits, expected outcomes, and alternatives. Three elements are required for truly informed consent, namely, competence, knowledge, and volition.

Institutional Review Board (IRB) A federally mandated oversight board that is required to monitor research and subject protections in institutions that engage in research. At a university all research proposals attain IRB approval before the project can be started to ensure that subject rights are protected, that federal regulations are followed, and documentation is provided (e.g., informed consent, reporting on adverse effects that arise in a study). The IRB also investigates allegations of violations of rights and provides reports of such investigations to the Department of Health and Human Services oversight commission.

Instrumentation A threat to internal validity that refers to changes in the measuring instrument or measurement procedures over time.

Intent-to-Treat Analysis. A way of handling missing data by replacing a missing value (e.g., on one or more measures at posttreatment or follow-up) with the last (previous) observation provided by the subject. An alternate name for this procedure is last-observation-carried-forward and describes how this is accomplished. The goal of the procedure is to retain rather than delete subjects and hence preserve the randomization of groups by keeping all subjects in the study. Contrast with Completer Analysis

Interaction Also called, statistical interaction. The combined effect of two or more variables as demonstrated in a factorial design. Interactions signify that the effect of one variable (e.g., sex of the subject) depends on the level of another variable (e.g., age).

Internal Consistency The degree of consistency or homogeneity of the items within a scale. Different reliability measures are used (e.g., Cronbach's alpha, split-half reliability, Kuder-Richardson 20 Formula).

Internal Validity The extent to which the experimental manipulation or intervention, rather than extraneous influences, can account for the results, changes, or group differences.

Interrater (or Interscorer) Reliability The extent to which different assessors, raters, or observers agree on the scores they provide when assessing, coding, or classifying subjects' performance.

Invasion of Privacy Seeking information of a personal nature that intrudes on what individuals or a group may view as private.

Investigator The person who is responsible for designing and planning the experiment.

Latin Square The arrangement of experimental conditions in a multiple-treatment design in which each of the conditions (task, treatments) occurs once in each ordinal position. Separate groups are used in the design, each of which receives a different sequence of the conditions.

Longitudinal Study Research that seeks to understand the course of change or differences over time by following (assessing) a group or groups over time, often involving several years. Contrast with Cross-Sectional Study.

Loose Protocol Effect A term to refer to the failure of the investigator to specify critical details of the procedures that guide the

experimenter's behavior, including the rationale, script, or activities of the investigation.

Magnitude of Effect A measure of the strength of the experimental effect or the magnitude of the contribution of the independent variable to performance on the dependent variable.

Main Effect The main effect is equivalent to an overall effect of an independent variable. In a factorial design, main effects are the separate and independent effects of the variables in the design, and are distinguished from interactions. See **Interaction**.

Masked A term sometimes used instead of "blind" to denote a procedure in which the experimenter and others associated with the investigation (e.g., staff, assessors) are kept naive with respect to the hypotheses and experimental conditions. The term "blind" is retained in this course because it continues to be the more frequent term and as a key word in searching resources on methodology. See **Blind**.

Matching Grouping subjects together on the basis of their similarity on a particular characteristic or set of characteristics that is known or presumed to be related to the independent or dependent variables.

Maturation Processes within the individual reflecting changes over time that may serve as a threat to internal validity.

Measurement Sensitivity The capacity of a measure to reflect systematic variation, change, or differences in response to an intervention, experimental manipulation, or group composition (e.g., as in a case control study).

Mechanism The steps or processes through which the intervention (or some independent variable) actually unfolds and exerts its influence. Mechanism explains more about underlying processes and how they lead change and goes beyond merely a statistical association (mediation).

Mediator A construct that shows a statistical relation between an experimental manipulation or intervention and the dependent variable or outcome. This is an intervening construct that suggests processes about why change occurs or on which change depends.

Meta-Analysis A quantitative method of evaluating a body of research in which effect size is used as the common metric. Studies are combined so that inferences can be drawn across studies and as a function of several of their characteristics (e.g., types of interventions).

Methodology The diverse principles, procedures, and practices that govern scientific research. In the present text, five components of methodology are distinguished: research design, assessment, data evaluation, ethical issues and responsibilities, and communication of findings.

Mismatching A procedure in which an effort is made to equalize groups that may be drawn from different samples. The danger is that the sample might be equal on a pretest measure of interest but regress toward different means upon retesting. Changes due to statistical regression might be misinterpreted as an effect due to the experimental manipulation.

Mixed-Method Research This is research that combines quantitative and qualitative research methods and occasionally is seen as a separate paradigm with its own literature, guidelines, and strategies. The importance of the area for this text is to convey that strategies that come from quite different traditions can be integrated and combined in given study.

Moderator A variable or characteristic that influences the direction or magnitude of the relation between two or more other variables (A and B). If the effect of an experimental manipulation varies as a function of some other characteristic (e.g., sex, ethnicity, temperament, genetics, neural activity), that other characteristic is referred to as a moderator.

Multigroup Cohort Design A prospective study in which two (or more) groups are identified at the initial assessment (time 1) and followed over time to examine outcomes of interest. One group is identified because they have an experience, condition, or characteristic of interest (exposure to domestic violence in the home); the other group is identified who does not have that experience.

Multiple Comparisons The number of comparisons or statistical tests in an experiment.

Multiple Operationism Defining a construct by several measures or in several ways. Typically, researchers are interested in a general construct (e.g., depression, anxiety) and seek relations among variables that are evident beyond any single operation or measure to define the construct.

Multiple-Baseline Design A single-case experimental design strategy in which the intervention is introduced across different behaviors, individuals, or situations at different points in time. A causal relation between the intervention and performance on the dependent measures is demonstrated if each behavior (individual or situation) changes when and only when the program is introduced.

Multiple-Treatment Design A design in which two or more different conditions or treatments are presented to each subject. In most multiple-treatment designs in clinical research, separate groups are used so that the different treatments (e.g., A, B) can be presented in different orders (A then B to one group and B then A to the other group).

Multiple-Treatment Interference A potential threat to external validity when subjects are exposed to more than one condition or treatment within an experiment. The impact of a treatment or intervention may depend on the prior conditions to which subjects were exposed.

Multitrait-Multimethod Matrix The set of correlations obtained from administering several measures to the same subjects. These measures include two or more constructs (traits or characteristics) each of which is measured by two or more methods (e.g., self-report, direct observation). The purpose of the matrix is to evaluate convergent and discriminant validity and to separate trait from method variance.

My Dissertation Committee A group of eminent scholars whose identity is completely protected because they entered the DCWPP immediately after my dissertation orals. (DCWPP stands for Dissertation Committee Witness Protection Program that provides a change of identity, relocation, and a gift certificate for plastic surgery. Wherever you are, thank you again for your help.)

N The overall sample size or number of subjects in a study and not to be confused with *n* which is the number of subjects in each of the groups.

Negative Results A term commonly used to refer to a pattern of experimental results in which the differences or findings are not statistically significant.

No-Contact Control Group A group that does not receive the experimental condition or intervention; subjects do not know they are participating in the research.

Nonequivalent Control Group A group used in quasi-experiments to rule out or make less plausible specific threats to internal validity. The group is referred to as nonequivalent because it is not formed through random assignment in the investigation.

Nonmanipulated Variables Variables that are studied through selection of subjects or observation of characteristics imposed by nature. See **Subject-Selection Study**.

Nonoverlapping Data In single-case designs, the data points from one phase (e.g., in an ABAB design) may not share any values so that there is no "overlap" in the graph when the data are plotted. This pattern is often evident when there are changes in means, slope, and level across phases, and the latency of change is rapid, all criteria that are used for nonstatistical evaluation of the data in single-case research.

Nonspecific Treatment Control Group See **Attention-Placebo Control Group**.

Nonstatistical Evaluation A method of data evaluation in single-case experimental research based on visual inspection criteria. Characteristics of the data (e.g., changes in means, slopes, and level, and the latency of change) are used to infer reliability of the impact of the experimental manipulation.

Normative Comparison A comparison of the individual with others, especially with a group of individuals who are functioning adequately in everyday life.

Normative Range A range of performance among a nonreferred, community sample that is used as a point of reference for evaluating the clinical significance of change in intervention studies.

No-Treatment Control Group A group that does not receive the experimental condition or intervention.

Novelty Effects A potential threat to external validity when the effects of an intervention may depend in part upon their innovativeness or novelty in the situation. The effects are genuine (i.e., nonchance), but occur because of the context in which they are studied. The same effect might not be evident when the intervention is part of routine or expected events, i.e., is not novel.

Nuisance Variables Characteristics of subjects (e.g., age, sex, ethnicity) that are not of interest to the investigator but that may vary systematically across groups and bias the results. In experimental research, random assignment of subjects to conditions or groups is a way of ensuring that such variables will be distributed unsystematically across groups. In this way, variables are not likely to threaten validity (e.g., by selection).

Null Hypothesis (H_0) The hypothesis that specifies that there is no difference between conditions or groups in the experiment on the dependent measures of interest.

Null Hypothesis Statistical Testing In the dominant model of research within the quantitative tradition, a study tests the null hypothesis, i.e., by posing that the experimental manipulation will have no effect. The null hypothesis is rejected if the differences between groups are statistically significant by a predetermined criterion (typically $p < .05$). Rejecting or accepting a hypothesis does not necessarily mean it is true or false (cf. Type I and Type II errors).

Objective Measures A class of assessment techniques that specify the items and response formats. The measures are fixed in the sense that the content and ways of answering are provided. Prime examples of objective measures are self- or other report scales of symptoms or daily functioning and questionnaires that measure ability, personality, and intelligence. The term "objective" has meaning in assessment in the context of "projective" measures in which stimuli and response format may be open ended.

Observational Research A type of research design in which the relations among variables are observed but not manipulated. Typically, the focus is on characteristics of different subjects or the relations among nonmanipulated variables.

Obtrusive Measures Any measure or measurement condition in which subjects are aware that some facet of their performance is assessed. See **Reactivity**.

Ongoing Assessment A feature of single-case experimentation in which observations of client functioning are obtained repeatedly (e.g., daily) over time.

Operational Definition Defining a concept by the specific operations or measures that are to be used in an experiment. The specific way the construct will be defined for inclusion in the investigation.

Order Effects In multiple-treatment designs, the impact of a treatment may depend on whether it appears first (or in some other place) among the treatments presented to the subjects. If the position of the treatments influences the results, this is referred to as an order effect. See also **Sequence Effects**.

Outlier An observation or score that departs greatly from rest of the scores in the data. The score is not merely at the high or low ranges but are conspicuously separated numerically from the next nearest scores and from the mean. There is no standard definition used but three or four standard deviations are sometimes used. Extreme scores can distort the overall distribution. Occasionally such scores are eliminated from the data, but the practice is not uniformly endorsed.

***p* level or value** A value associated with the statistical test (e.g., *t* or *F* test) that reflects the probability that a value as or more extreme than the one observed would arise by chance alone, if the study were repeated a large number of times.

Parsimony An accepted principle or heuristic in science that guides our interpretations of data and phenomena of interest. The principle refers to selecting the simplest version or account of the data among the competing views or interpretations that are available. If a phenomenon can be explained equally well in multiple ways, one adopts the interpretation that is most economical, i.e., uses the fewest constructs. Other names of the principle include the principle of economy, principle of unnecessary plurality, principle of simplicity, and Occam's razor.

Participant The person who is the subject or who takes part and provides the data for the study. This is used interchangeably with subject, although in research with humans (rather than nonhuman animals), participant tends to be the preferred term.

Patched-up Control Group See **Nonequivalent Control Group**.

Per-Comparison Error Rate The probability of a Type I error for a specific comparison or statistical test of differences when several comparisons are made. Contrast with Experiment-Wise Error Rate.

Physical Traces Unobtrusive measures that consist of selective wear (erosion) or the deposit (accretion) of materials.

Pilot Work A preliminary test of the procedures of an investigation before running the full-fledged study. Usually, the goals of pilot work are to see if procedures (e.g., equipment, recruitment methods) "work," are feasible, and are having the effect (e.g., on the manipulation check or even dependent measures). Pilot work usually is conducted in a small scale merely to provide the information the investigator wishes to assure that the study can be conducted.

Placebo Effect Change in an outcome due to expectancies for improvement on the part of clients or those who are administering a medication or other intervention procedure.

Placebo A substance that has no active pharmacological properties that would be expected to produce change in the condition to which it is applied.

Plagiarism Refers to the direct use and copying of material of someone else without providing credit or acknowledgment. This can include words or ideas that from another person that one attributes to oneself.

Plausible Rival Hypothesis An explanation of the results of an investigation that is reasonable and includes other influences than the one the investigator has studied. One or more of the many threats to internal, external, construct, and data-evaluation validity may be a plausible rival hypothesis.

Postexperimental Inquiry A method of evaluating whether demand characteristics could account for the results by asking the subjects after the experiment about their perceptions of the purpose of the experiment, what the experimenter expected from them, and how they were supposed to respond.

Posttest Sensitization Administration of a measure after an experimental manipulation might crystalize the reactions of participants and influence performance. If participants can connect the measure to the prior experience, sensitization is more likely.

Posttest-Only Control Group Design An experimental design (with a minimum of two groups) in which no pretest is given.

In a true-experimental version, participants are assigned randomly to conditions. The effect of the experimental condition between or among groups is assessed on a postintervention measure only.

Power The probability of rejecting the null hypothesis (that there are no differences) when in fact that hypothesis is false, i.e., there are no differences in fact. That is, correctly rejecting the null hypothesis.

Practical Significance A term used in the context of applied research (e.g., clinical and counseling psychology, education, medicine, business, and industry). There is no standard index or measure of practical significance, so the term is used loosely usually to ask whether a particular finding would make any "real" difference in everyday life. In clinical psychology, in the contest of treatment studies clinical significance is the term that is used instead and has a number of commonly used indices.

Predictive Validity The correlation of a measure at one point in time with performance on another measure or criterion at some point in the future.

Preinquiry A method of evaluating whether demand characteristics could account for the results by conveying information to the subjects about the experiment without actually running them through the conditions. Subjects are also asked to complete the dependent measures to see if their performance yields the expected results.

Pretest Sensitization Administration of the pretest may alter the influence of the experimental condition that follows.

Pretest-Posttest Control Group Design An experimental design with a minimum of two groups. Usually, one group receives the experimental condition and the other does not. In the true experimental version, participants are assigned randomly to conditions. The essential feature of the design is that subjects are tested before and after the intervention.

Probability Pyramiding The error rate or risk of a Type I error rate that comes from conducting multiple comparisons (e.g., *t* tests) in an experiment.

Projective Measures A class of assessments techniques that attempt to reveal underlying motives, processes, styles, themes, personality, and other psychological process. These characteristics are measured indirectly. Clients are provided with an ambiguous task where they are free to respond with minimal situational cues or constraints. The ambiguity of the cues and minimization of stimulus material allow the client to freely "project" onto the situation important processes within his or her own personality.

Proof of Concept This is a demonstration to show that something can occur. This may be a demonstration in a situation that is artificial, contrived, and in a laboratory context that does not mimic the world in ever day lie. The goal is to show whether something can happen and not whether it does in fact occur that way in everyday life.

Propensity Score Matching A set of statistical procedures that integrate multiple variables that may influence selection when groups are compared on a particular outcome. The goal is to construct groups that are matched on a large set that contributed to group selection, i.e., those variables for whatever reason led some subjects to be in one condition or group rather than the other group. The matching makes less plausible the impact of differences due to variables other than the intervention.

Prospective Study A design in which one or more samples are followed over time. Initial assessment or evaluation of a characteristic of the sample is related to some outcome at a future point in time.

Protective Factor A variable that prevents or reduces the likelihood of a deleterious outcome. The concept usually is invoked in the context of identifying a special populations or group that is at risk for a particular outcome (e.g., a clinical disorder, delinquency, drug use). A protective is any variable associated with reduction in that risk. This is a correlation of some characteristic that reduces the likelihood of the deleterious outcome.

Psychobiological Measures Refer to assessment techniques designed to examine biological substrates and correlates of affect, cognition, and behavior and the links between biological processes and psychological constructs. The measures encompass many different types of functions (e.g., arousal of the autonomic system), systems (e.g., cardiovascular, gastrointestinal, neurological), and levels of analysis (e.g., microelectrode physiology that permits analysis of the response of individual neurons in the brain and brain imaging in response to tasks and activities in human and nonhuman animal research).

Psychometric Characteristics A general term that encompasses diverse types of reliability and validity evidence in behalf of a measure.

Publication Bias When manuscripts are considered for publication in scientific journals, those manuscripts with findings that are statistically significant, so-called positive results, are much more likely to be published than those with findings that are not statistically significant, so-called negative effects.

Qualitative Research An approach to research that focuses on narrative accounts, description, interpretation, context, and meaning. The goal is to describe, interpret, and understand the phenomena of interest and to do so in the context in which experience occurs. The approach is distinguished from the more familiar Quantitative Research.

Quantitative Research The dominant paradigm for empirical research in psychology and the sciences more generally involving the use of operational definitions, careful control of the subject matter, efforts to isolate variables of interest, quantification of constructs, and null hypothesis and statistical testing. This is distinguished from Qualitative Research.

Quasi-Experimental Design A type of design in which the conditions of true experiments are only approximated. Restrictions are placed on some facet of the design such as the assignment of cases randomly to conditions and that affects the strength of the inferences that can be drawn.

Random Assignment Allocating or assigning subjects to groups in such a way that the probability of each subject appearing in any of the groups is equal. This usually is accomplished by determining in the group to which each subject is assigned by an online program that provides sets of numbers that correspond to the number of groups or conditions and place them in a random order or by looking at a table of random numbers and going in order across the rows and/or columns and pulling out the needed numbers in the order they appear in the table.

Random Selection Drawing subjects from a population in such a way that each member of the population has an equal probability of being drawn.

Randomized Controlled Trial A treatment outcome study in which clients with a particular problem (e.g., depression, cancer) are randomly assigned to various treatment and control conditions. This is a type of true-experiment (usually a pretest–posttest control group design) and is regarded by many as the "gold standard," i.e., the best and most definitive way of demonstrating that an intervention is effective.

Reactivity Performance that is altered as a function of subject awareness (e.g., of the measurement procedures, of participation in an experiment).

Recovery A concept used in the context of evaluating improvements in mental disorders and addictions. The focus is on improvements in different spheres of functioning (e.g., health, stable living conditions, having a purpose, being involved in relationships and health). The purpose of the construct is to define meaningful adjustment and participation in life rather than the mere absence or reduction of symptoms.

Regression Effect See **Statistical Regression**.

Reliability Refers to consistency of the scores obtained for a measure. This can encompasses consistency in different ways, including among items of the measure (i.e., how the items relate to each other), consistency between different parts or alternate forms of the same measure, and consistency in performance on the measure over time (test–retest for a given group of subjects).

Reliability of Change Index Refers to a measure to evaluate clinical significance of change from pretreatment to posttreatment in the context of therapy or another intervention. A commonly used criterion is a change (improvement) of a client's score at posttreatment that is at least 1.96 standard deviations better than (departs from) the pretreatment mean for the group. The criterion (1.96) was adopted because this is used in a different context (e.g., *t* tests to determine whether two groups are statistically different from each other at the $p < .05$ level).

Replication Repetition of an experiment or repetition of the findings of an experiment.

Research Design The plan or arrangement that is used to examine the question of interest; the manner in which conditions are planned so as to permit valid inferences.

Response Set or Style In measurement refers to a systematic way of answering questions or responding to the measure that is separate from the construct of interest. Socially desirable responding and acquiescence are two examples. In each case, participants will answer in keeping with the response set (e.g., placing themselves in a good light).

Response Shift Changes in a person's internal standards of measurement. The shift reflects a change in values, perspective, or criteria that lead to evaluation of the same or similar situations, behaviors, states, in a different way. The threshold or standards a person invokes have changed although the actual instrument or measure remains the same. This can be a special case of instrumentation as a threat to internal validity.

Retrospective Design A case-control design draws inferences about some antecedent condition that has resulted in or is associated with the outcome. Subjects are identified who already show the outcome of interest (cases) and are compared with those who do not show the outcome (controls). Assessment focuses on some other characteristic in the past.

Retrospective Study A design in which individuals are assessed on a characteristic of interest and as well recount event or experiences in the past. All the assessments are done in the present, but the goal is to identify what might have occurred earlier in life to predict or explain the present outcome.

Reversal Phase A phase or period in single-case designs in which the baseline (nonintervention) condition is reintroduced to see if performance returns to or approximates the level of the original baseline.

Risk Factor A characteristic that is an antecedent to and increases the likelihood of an outcome of interest. A "correlate" of an outcome of interest in which the time sequence is established.

Sample Size The number of subjects or cases included in a study. This can refer to the overall number of subjects in the study (N) or the number of subjects within a group (n).

Samples of Convenience Subjects included in an investigation who appear to be selected merely because they are available, whether or not they provide a suitable or optimal test of the hypotheses or conditions of interest.

Secondary Data Analyses Refers to conducting empirical studies based on data already collected and available. That is, one does not "run" subjects in the sense of collecting new data, but rather draws on available data sets. Meta-analysis is one example of this type of analysis.

Self-plagiarism Refers to presentation of one's own prior work (material, quotes, ideas) without acknowledgment and passing off the material as if it is new. Variations include: submitting a published paper to a second outlet (duplicate publication), copying select sections of text or figures and publishing those, copying from one's prior work, presenting the same data again as if they were not presented previously.

Self-Report Inventories Questionnaires and scales in which the subjects report on some facet of their functioning (e.g., personality, cognitions, opinions, behaviors).

Sensitivity When we are interested in predicting an outcome (e.g., who will be a terrorist, who will get a particular disease), we use variables that relate to the outcome. Sensitivity refers to the rate or probability of identifying individuals who are predicted to show a particular characteristic variables and in fact do show that predicted outcome. These are also called true positives. For example, the probability of being a heavy cigarette smoker (predictor) and in fact later having lung disease (outcome) would be sensitivity.

Sequence Effects In multiple-treatment designs, several treatments may be presented to the subject. A series of treatments is provided (e.g., treatments A, B, then C for some subjects and B, C, then A for other subjects, and so on for other combinations). If the sequence yields different outcomes, this is referred to as sequence effects. See **Order Effects**.

Significance Fallacy Refers to the interpretation of "statistical significance" as being a measure of real or important differences (e.g., practical or clinically significant differences). The fallacy is that there is no necessary relation between statistical and clinical significance or what is "significant" statistically is not necessarily "significant" or important in any other way.

Significance Level See **Alpha**.

Simulators A method of estimating whether demand characteristics could explain the findings. Subjects are asked to act as if they received the treatment or intervention even though they actually do not. These simulators are then run through the assessment procedures of the investigation by an experimenter who is "blind" as to who is a simulator and who is a real subject.

Single Operationism Defining a construct by a single measure or one operation. Contrast with Multiple Operationism.

Single-Case Experimental Designs Research designs in which the effects of an intervention can be evaluated with the single case, i.e., one subject. The designs can be used for multiple-cases and groups and are distinguished by several features such as ongoing assessment of participants over time and drawing inferences from repeated changes in performance as a function of altering conditions as the various designs (e.g., ABAB, multiple baseline) dictate.

Social Impact Measures Measures in outcome research that are important in everyday life or to society at large (e.g., truancy, arrest records, utilization of health services).

Socially Desirable Response Set This is a way of responding to a measure so as to place oneself in a socially desirable light. This response set or style of responding can compete with obtaining an individual's true score on the construct of interest in the measure.

Solomon Four-Group Design An experimental design that is used to evaluate the effect of pretesting. The design can be considered as a combination of the pretest–posttest control group design and a posttest-only design in which pretest (provided vs. not provided) and the experimental intervention (treatment vs. no treatment) are combined.

Specificity When we are interested in predicting an outcome (e.g., who will be a terrorist, who will get a particular disease), we use variables that relate to the outcome on some assessment. As part of the prediction, some individuals do not show the early predictors

of some outcome and in fact do not show the outcome later. This is the rate or probability of identifying individuals who are *not* likely to show an outcome and in fact do not. These also are called true negatives. For example, the probability of not being a cigarette smoker early in life and also not having lung disease later would be specificity.

Stable Rate Performance obtained from ongoing observations over time, as in single-case experimental designs, in which there is little or no variability in the data.

Statistical Evaluation Applying statistical tests to assess whether the obtained results are reliable or can be considered to be sufficient to reject the null hypothesis.

Statistical Power See **Power**.

Statistical Regression The tendency of extreme scores on any measure to revert (or regress) toward the mean of a distribution when the measure is administered a second time. Regression is a function of the amount of error in the measure and the test–retest correlation.

Statistical Significance A criterion used to evaluate the extent to which the results of a study (e.g., differences between groups or changes within groups) are likely to be due to genuine rather than chance effects. A statistically significant difference indicates that the probability level is equal to or below the level of confidence selected (e.g., $p < .05$), i.e., if the experiment were conducted repeatedly, the finding would occur 5/100 times on a chance basis.

Subject See **Participant**.

Subject Variables Those variables that are based on features within the individual or circumstances to which they were exposed. These variables usually are not manipulated experimentally.

Subjective Evaluation A method of evaluating the clinical significance of an intervention outcome by assessing the opinions of clients themselves, individuals who are likely to have contact with the client, or persons in a position of expertise. The question addressed by this method of evaluation is whether changes in treatment have led to differences in how the client is viewed by others or how the client views herself or himself.

Subject-Selection Biases Factors that operate in selection of subjects or selective loss or retention of subjects over the course of the experiment that can affect experimental validity. Primary examples would be selection, recruitment, or screening procedures that might restrict the generality (external validity) of the findings and loss of subjects (attrition) that might alter group composition and lead to differences that would be mistaken for an intervention effect (internal validity).

Systematic Replication A study designed to repeat a prior experiment but allows features of the original study to vary. The conditions and procedures of the replication are deliberately designed only to approximate those of the original experiment.

Test Sensitization Alteration of subject performance due to administration of a test before (pretest) or after (posttest) the experimental condition or intervention. The test may influence (increase, decrease, nullify) the effect of the experimental condition. A potential threat to external validity if the effect of the experimental condition may not generalize to different testing conditions.

Testing A threat to internal validity that consists of the effects of taking a test on repeated occasions. Performance may change as a function of repeated exposure to the measure rather than to the independent variable or experimental condition.

Test–Retest Reliability The stability of test scores over time; the correlation of scores from one administration of the test with scores on the same instrument after a particular time interval has elapsed.

Theory The conceptualization of the phenomena we are studying. This is an explanation of how variables relate to each other, how are they connected, and the implications can we draw from that for research. A theory is designed to explain but also to generate hypotheses that can be used to test or revise the theory.

Threat to Validity This is a potential influence in the study that will interfere with drawing valid (accurate, sound) inferences about the effect of the experimental manipulation or intervention. Threats can result from a variety of factors included under the broader rubrics of internal, external, construct, and data-evaluation validity.

Transferability A criterion invoked to evaluate data in qualitative research and pertains to whether the data are limited to particular context (are context bound) and is evaluated by looking at any special characteristics (unrepresentativeness) of the sample.

Translational Research Refers to research that moves from a basic finding to application. The full process can be characterized as moving a finding from bench (basic, laboratory research) to bedside (clinical application with patients), or the community (large-scale application if pertinent to public health).

Transparency A core value of science reflects openness about what one is doing and has done in research. One provides information (e.g., methods, procedures, data) to others whenever possible so that the work can be scrutinized and that peers can view and replicate the findings. Transparency also includes public access to research methods as well.

Treatment as Usual The routine treatment that is provided in a given setting for the same clinical problem or intervention focus. This group receives whatever is usually done, i.e., as usual care.

Treatment Differentiation The demonstration showing that two or more treatments were distinct along predicted dimensions. This complements but is distinguishable from treatment integrity.

Treatment Integrity The fidelity with which a particular treatment is rendered in an investigation. Integrity includes whether the treatment was provided and provided as intended.

Triangulation The extent to which data from separate sources converge to support the conclusions. Also used as a criterion to evaluate data in qualitative research.

True Experiment A type of research in which the arrangement permits maximum control over the independent variables or conditions of interest. The investigator is able to assign subjects to different conditions on a random basis, to include alternative conditions (e.g., treatment and control conditions) as required by the design, and to control possible sources of bias within the experiment that permit the comparison of interest.

Trustworthiness A criterion used to evaluate data in qualitative research. The criterion includes multiple components, namely, credibility, transferability, dependability, and confirmability of the data.

Type I Error See **Alpha**.

Type II Error See **Beta**.

Unobtrusive Measures Those measures that are outside of the awareness of the subject.

Validity Refers to the content of a measure and whether the measure assesses the domain of interest. This encompasses the relation of performance on the measure to performance on other measures at the same time or in the future and to other criteria with which the scores on the measure would be expected to be related (e.g., school achievement, occupational status, psychiatric diagnosis).

Visual Inspection A method of data evaluation commonly used in single-case research based on examining the pattern of change (means, level, slope, latency of change) over phases.

Volition A requirement of informed consent is an assurance that the subject agrees to participate without coercion. For subjects to

provide consent, they must have a choice to participate or not and to withdraw or change their minds later even after they have provided consent.

Wait-List Control Group A group that is designed to control for threats to internal validity. The experimental condition or intervention is not provided during the period that experimental subjects receive the intervention. After this no-treatment period, subjects in this control group receive the intervention.

WEIRD This is an acronym for Western, Educated, Industrialized, Rich, and from Democratic Cultures. The term was coined to note that much research in the United States relies on undergraduate students who are WEIRDos, and do not necessarily represent individuals from other cultures in fundamental ways. This is an issue of external validity, i.e., do the results obtain with college student samples generalize—many findings do not.

Yoking Matching subjects in different groups on some variable (e.g., duration or number of sessions) that might emerge during the course of the study. The investigator wishes to rule out the impact of these probably ancillary differences between intervention and non-intervention groups. Subjects in intervention and nonintervention groups are paired as "partners" so to speak and the emergent variable (e.g., more sessions) that were provided to the intervention subject is assigned to the partner. Not necessarily a separate group from one of the prior groups in this table.

References

Aas, I. M. (2010). Review Global Assessment of Functioning (GAF): Properties and frontier of current knowledge. *Annals of General Psychiatry*, *9*, 20. Available at www.biomedcentral.com/content/pdf/1744-859X-9-20.pdf

Aas, I. M. (2011). Guidelines for rating global assessment of functioning (GAF). *Annals of General Psychiatry*, *10*(2), 1–11.

Abdullah, M. M., Ly, A. R., Goldberg, W. A., Clarke-Stewart, K. A., Dudgeon, J. V., Mull, C. G., . . . Ericson, J. E. (2012). Heavy metal in children's tooth enamel: Related to autism and disruptive behaviors? *Journal of Autism and Developmental Disorders*, *42*(6), 929–936.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*(3), 361–382.

Adam, Y., Meinlschmidt, G., & Lieb, R. (2013). Associations between mental disorders and the common cold in adults: A population-based cross-sectional study. *Journal of Psychosomatic Research*, *74*(1), 69–73.

Adams, J. B., Baral, M., Geis, E., Mitchell, J., Ingram, J., Hensley, A., . . . El-Dahr, J. M. (2009). The severity of autism is associated with toxic metal body burden and red blood cell glutathione levels. *Journal of Toxicology*. Available at www.hindawi.com/journals/jt/2009/532640/

Adeponle, A. B., Thombs, B. D., Groleau, D., Jarvis, E., & Kirmayer, L. J. (2012). Using the cultural formulation to resolve uncertainty in diagnoses of psychosis among ethnoculturally diverse patients. *Psychiatric Services*, *63*(2), 147–153.

Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(1), 93–101.

Adler, N., Bush, N. R., & Pantell, M. S. (2012). Rigor, vigor, and the study of health disparities. *Proceedings of the National Academy of Sciences*, *109*(Supplement 2), 17154–17159.

Agency for Health Care Policy and Research. (1999). *Treatment of depression—Newer pharmacotherapies* (Publication No. 99-E014). Evidence Report/Technology Assessment No. 7, Rockville, MD.

Ahearn, W. H., Clark, K. M., MacDonald, R. P. F., & Chung, B. I. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis, 40*, 263–275.

Alberts, B. (2013). Editorial: Impact factor distortions. *Science, 340*, 787.

Al-Farsi, Y. M., Waly, M. I., Al-Sharbati, M. M., Al-Shafaee, M. A., Al-Farsi, O. A., Al-Khaduri, M. M., . . . Deth, R. C. (2013). Levels of heavy metals and essential minerals in hair samples of children with autism in Oman: A case–control study. *Biological Trace Element Research*, *151*(2), 181–186.

Ali, A., Hossain, M., Hovsepian, K., Rahman, M., Plarre, K., & Kumar, S. (2012). mPuff: Automated detection of cigarette smoking puffs from respiration measurements. In *Proceedings of ACM.* IPSN, Beijing, China.

Allen, A. P., & Smith, A. P. (2012). Demand characteristics, pre-test attitudes and time-on-task trends in the effects of chewing gum on attention and reported mood in healthy volunteers. *Appetite*, *59*(2), 349–356.

Allen, J., Weinrich, M., Hoppitt, W., & Rendell, L. (2013). Network-based diffusion analysis reveals cultural transmission of lobtail feeding in humpback whales. *Science*, *340*(6131), 485–488.

Allen, K. D., & Evans, J. H. (2001). Exposure-based treatment to control excessive blood glucose monitoring. *Journal of Applied Behavior Analysis*, *34*, 497–500.

Al-Sahab, B., Heifetz, M., Tamim, H., Bohr, Y., & Connolly, J. (2012). Prevalence and characteristics of teen motherhood in Canada. *Maternal and Child Health Journal*, *16*(1), 228–234.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS One*, *6*(9), e2435.

Alvarez, M. M. R., Fernández, M. M. M., Conroy, B. V., & Martínez, A. C. (2008). Criteria of the peer review process for publication of experimental and quasi-experimental research in Psychology: A guide for creating research papers. *International Journal of Clinical and Health Psychology*, *8*(3), 751–764.

American Academy of Pediatrics. (2012, February). AAP Reaffirms Breastfeeding Guidelines. Retrieved June 2013 from www.aap.org/en-us/about-the-aap/aap-press-room/pages/AAP-Reaffirms-Breastfeeding-Guidelines.aspx

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed. Revised). Washington, DC: Author.

American Psychiatric Association. (2000). Practice guideline for the treatment of patients with major depressive disorder (revision). *American Journal of Psychiatry*, *157* (supplement 4), 1–45.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association. (2001a). Ethical principles of psychologists and code of conduct: Draft for comment. *Monitor in psychology*, *32*(2), 77–89.

American Psychological Association. (2002). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. Washington, DC: American Psychological Association. Available at www.apa.org/pi/oema/resources/policy/multicultural-guidelines.aspx

American Psychological Association. (2008). Publications and Communications Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851.

American Psychological Association. (2010a). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association. Available at www.apa.org/ethics/code/principles.pdf

American Psychological Association. (2010b). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

American Psychological Association. (2011). Journals by title. Available at http://search.apa.org/publications?query=&facet=&pubtype=journals&section=title&sort=titleBa

American Psychological Association. (2014). *Journals program of the American Psychological Association.* Washington, DC: Author.

American Psychological Association. (2001b). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, C. A., Benjamin, A. J., & Bartholow, B. D. (1998). Does the gun pull the trigger? Automatic priming effects of weapon pictures and weapon names. *Psychological Science*, *9*(4), 308–314.

Anderson, G. M., Jacobs-Stannard, A., Chawarska, K., Volkmar, F. R., & Kliman, H. J. (2007). Placental trophoblast inclusions in autism spectrum disorder. *Biological Psychiatry*, *61*(4), 487–491.

Andresen, R., Caputi, P., & Oades, L. (2006). Stages of recovery instrument: Development of a measure of recovery from serious mental illness. *Australian and New Zealand Journal of Psychiatry*, *40*(11–12), 972–980.

Aneshensel, C. S., & Stone, J. D. (1982). Stress and depression: A test of the buffering model of social support. *Archives of General Psychiatry, 39,* 1392–1396.

Aneshensel, C. S., Phelan, J. C., & Bierman, A. (Eds.) (2013). *Handbook of the sociology of mental health* (2nd ed.). Dordrecht, the Netherlands : Springer Science + Business Media.

Anholt, R. R., & Mackay, T. F. (2012). Genetics of aggression. *Annual Review of Genetics*, *46*, 145–164.

Arguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16,* 270–301.

Arguinis, H., & Vandenberg, R. J. (Eds.) (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*, 569–595.

Arndt, J. E., Hoglund, W. L., & Fujiwara, E. (2013). Desirable responding mediates the relationship between emotion regulation and anxiety. *Personality and Individual Differences*, *55*(2), 147–151.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Atran, S., Medin, D. L., &. Ross, N. O. (2005). The cultural mind: environmental decision making and cultural modeling within and across populations. *Psychological Review*, *112*, 744–776.

Aureli, F., Schaffner, C. M., Verpooten, J., Slater, K., & Ramos-Fernandez, G. (2006). Raiding parties of male spider monkeys: Insights into human warfare? *American Journal of Physical Anthropology*, *131*(4), 486–497.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424 (Special journal issue on Propensity Score Analysis).

Awad, A. G., & Voruganti, L. N. (2012). Measuring quality of life in patients with schizophrenia. *Pharmacoeconomics*, *30*(3), 183–195.

Aylward, B., & Yamada, T. (2011). The polio endgame. *New England Journal of Medicine*, *364*(24), 2273–2275.

Aziz, S., Uhrich, B., Wuensch, K. L., & Swords, B. (2013). The Workaholism Analysis Questionnaire: Emphasizing work-life imbalance and addiction in the measurement of workaholism. Institute of Behavioral and Applied Management. Available at www.ibam.com/pubs/jbam/articles/vol14/No2/Article1.pdf

Babaria, P., Abedin, S., Berg, D., & Nunez-Smith, M. (2012). "I'm too used to it": A longitudinal qualitative study of third year female medical students' experiences of gendered encounters in medical education. *Social Science & Medicine*, *74*(7), 1013–1020.

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, *161*(12), 2163–2177.

Bakeman, R., & Quera, R. (2012). Behavioral observations. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 207–225). Washington, DC: American Psychological Association.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Balaji, M., Chatterjee, S., Koschorke, M., Rangaswamy, T., Chavan, A., Dabholkar, H., . . . Patel, V. (2012). The development of a lay health worker delivered collaborative community based intervention for people with schizophrenia in India. *BioMed Central Health Services Research*, *12*. Available at www.biomedcentral.com/1472-6963/12/42/

Bandettini, P. A. (2012). Twenty years of functional MRI: The science and the stories. *NeuroImage*, *62*(2), 575–588.

Bang, M., Medin, D. L., & Atran, S. (2007). Cultural mosaics and mental models of nature. *Proceedings of the National Academy of Sciences*. Retrieved August 22, 2007, from 10.1073/pnas.0706627104.

Banister, P., Bunn, G., Burman, E., Daniels, J., Duckett, P., Goodley, D., . . . Whelan, P. (2011). *Qualitative methods in psychology: A research guide* (2nd ed.). Berkshire, England: Open University Press.

Bansal, G., Zahedi, F., & Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems*, *49*(2), 138–150.

Banton, M. (2010). The vertical and horizontal dimensions of the word race. *Ethnicities*, *10*, 127–140.

Barber, J. P., Barrett, M. S., Gallop, R., Rynn, M. A., & Rickels, K. (2012). Short-term dynamic psychotherapy versus pharmacotherapy for major depressive disorder: A randomized, placebo-controlled trial. *Journal of Clinical Psychiatry*, *73*(1), 66–73.

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. Elmsford, NY: Pergamon.

Bardo, M. T., & Pentz, M. A. (2012). Translational research. In H. Cooper, H. (Ed.), *APA handbook of research methods in psychology* (Vol. 3, pp. 553–568). Washington, DC: American Psychological Association.

Bargh, J. A. (Ed.) (2007). *Social psychology and the unconscious: The automaticity of higher mental processes*. New York: Psychology Press.

Bargh, J. A., & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, *3*, 73–79.

Bargh, J. A., Schwader, K. L, Hailey, S. E., Dyer, R. L., & Boothby, E. J. (2012). Automaticity in social-cognitive processes. *Trends in Cognitive Science, 16,* 593–605.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*, 12–18.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Chichester, UK: John Wiley & Sons.

Bartholow, B. D., Anderson, C. A., Carnagey, N. L., & Benjamin Jr., A. J. (2005). Interactive effects of life experience and situational cues on aggression: The weapons priming effect in hunters and nonhunters. *Journal of Experimental Social Psychology*, *41*(1), 48–60.

Baserga, R. (2011). Self-plagiarism in music and science. *Nature*, *470*(7332), 39–39.

Baskin, T. W., Tierney, S. C., Minami, T., & Wampold, B. E. (2003). Establishing specificity in psychotherapy: A meta-analysis of structural equivalence of placebo controls. *Journal of Consulting and Clinical Psychology*, *71*, 973–979.

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, A., . . . the STOPIT-2 Study Group (2010). Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *Journal of the American Medical Association, 303,* 1180–1187.

Batalla, A., Crippa, J. A., Busatto, G. F., Guimaraes, F. S., Zuardi, A. W., Valverde, O., . . . Martin-Santos, R. (2013). Neuroimaging studies of acute effects of THC and CBD in humans and animals: A systematic review. *Current Pharmaceutical Design*. Available at http://europepmc.org/abstract/MED/23829359/reload=0;jsessionid=yCqO2kAxMYihkk4NJtgL.2

Bateman, A., & Fonagy, P. (2010). Mentalization based treatment for borderline personality disorder. *World Psychiatry*, *9*(1), 11–15.

Bates, T., Anić, A., Marušić, M., & Marušić, A. (2004). Authorship criteria and disclosure of contributions. *Journal of the American Medical Association*, *292*(1), 86–88.

Baumann, J., & DeSteno, D. (2010). Emotion guided threat detection: Expecting guns where there are none. *Journal of Personality and Social Psychology*, *99*(4), 595–610.

Beall, A. T., & Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological science*, *24*(9), 1837–1841.

Beall, J. (2012, December). Criteria for determining predatory open-access publishers (2nd ed.). Available at www.sjsm.org/library/criteria-2012-2.pdf

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, *42*, 861–865.

Beecher, H. K. (1966). Ethics and clinical research. *The New England Journal of Medicine*, *274*, 1354–1360.

Behrens, J. T., DiCerbo, K. E., Yei, N., & Levy, R. (2013). Exploratory data analysis. In I.B. Weiner, J.A. Schinka, & W.F. Veliver (Eds.), *Handbook of psychology: Volume 2: Research methods in psychology* (2nd ed., pp. 34–70). New York: John Wiley & Sons.

Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. *Handbook of psychology*. Published online, John Wiley & Sons.

Belin, R. J., Greenland, P., Martin, L., Oberman, A., Tinker, L., Robinson, J., . . . Lloyd-Jones, D. (2011). Fish intake and the risk of incident heart failure: Clinical perspective The Women's Health Initiative. *Circulation: Heart Failure*, *4*(4), 404–413.

Benedetti, F. (2009). *Placebo effects: Understanding the mechanisms in health and disease*. New York: Oxford University Press.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine 8*(8): e1001083.

Benos, D. J., Fabres, J., Farmer, J., Gutierrez, J. P., Hennessy, K., Kosek, D., . . . Wang K. (2005). Ethics and scientific publication. *Advances in Physiology Education*, *29*, 59–74.

Bent-Hansen, J., & Bech, P. (2011). Validity of the definite and semidefinite Questionnaire version of the Hamilton Depression Scale, the Hamilton Subscale and the Melancholia Scale. Part I. *European Archives of Psychiatry and Clinical Neuroscience*, *261*(1), 37–46.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–542.

Bernasconi, B. (2010). Defining race scientifically: A response to Michael Banton. *Ethnicities*, *10*, 141–148.

Berns, K. I., Casadevall, A., Cohen, M. L., Ehrlich, S. A., Enquist, L. W., Fitch, J. P., . . . Vidaver, A. K. (2012). Adaptations of avian flu virus are a cause for concern. *Science*, *335*(6069), 660–661.

Beskow, L. M., Dame, L., & Costello, E. J. (2008). Certificates of confidentiality and the compelled disclosure of research data. *Science, 322,* 1054–1055.

Bhar, S. S., & Beck, A. T. (2009). Treatment integrity of studies that compare short-term psychodynamic psychotherapy with cognitive-behavior therapy. *Clinical Psychology: Science and Practice*, 16(3), 370–378.

Bickel, W. K., Yi, R., Landes, R. D., Hill, P. F., & Baxter, C. (2011). Remember the future: Working memory training decreases delay discounting among stimulant addicts. *Biological Psychiatry*, *69*(3), 260–265.

Biemer, P. P., Groves, R. W., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.) (2004). *Measurement errors in surveys*. Hoboken, NJ: John Wiley & Sons.

Bisakha, S. (2010). The relationship between frequency of family dinner and adolescent problem behaviors after adjusting for other family characteristics. *Journal of Adolescence, 33(1*), 187–196.

Blackless, M., Charuvastra, A., Derryck, A., Fausto-Sterling, A., Lauzanne, K., & Lee, E. (2000). How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology*, *12*, 151–166.

Blair, E. (2004). Discussion: Gold is not always good enough: The shortcomings of randomization when evaluating interventions in small heterogeneous samples. *Journal of Clinical Epidemiology*, *57*, 1219–1222.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78–84.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Blumenthal, J. A., Babyak, M. A., Doraiswamy, P. M., Watkins, L., Hoffman, B. M., Barbour, K. A., . . . Sherwood, A. (2007). Exercise and pharmacotherapy in the treatment of major depressive disorder. *Psychosomatic Medicine*, *69*(7), 587–596.

Bohannon, J. (2013).Who's afraid of peer review? *Science*, *342*, 60–65.

Bollier, D. (2010). *The promise and peril of big data*. Queenstown, MD: The Aspen Institute. Available at http://india.emc.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

Bolton, J. L., Huff, N. C., Smith, S. H., Mason, S. N., Foster, W. M., Auten, R. L., & Bilbo, S. D. (2013). Maternal stress and effects of prenatal air pollution on offspring mental health outcomes in mice. *Environmental Health Perspectives*, *121*(9), 1075–1082.

Bombardier, C. H., Ehde, D. M., Gibbons, L. E., Wadhwani, R., Sullivan, M. D., Rosenberg, D. E., & Kraft, G. H. (2013). Telephone-based physical activity counseling for major depression in people with multiple sclerosis. *Journal of Consulting and Clinical Psychology*, *81*(1), 89–99.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, *8*(4), 445–454.

Booth, B. M., Blow, F. C., & Cook, C. A. L. (1998). Functional impairment and co-occurring psychiatric disorders in medically hospitalized men. *Archives of Internal Medicine*, *158*, 1551–1559.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist*, *63*, 77–95.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, England: John Wiley & Sons.

Borsboom, D., & Wagenmakers, E. (2013). Derailed: The rise and fall of Diederik Stapel. *Observer*, *26*(1), 31, 33.

Bourgeois, F. T., Murthy, S., & Mandl, K. D. (2010). Outcome reporting among drug trials registered in Clinicaltrials.gov. *Annals of Internal Medicine*, *153*(3), 158–166.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time-series analysis: Forecasting and control* (3rd ed.). New Jersey: Prentice-Hall.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal*, *5*, 437–474.

Bradshaw, C. P., Debnam, K. J., Martin, L., & Gill, R. (2006). Assessing rates and characteristics of bullying through an internet-based survey system. *Proceedings of Persistently Safe Schools*, 147–157.

Bradshaw, W. (2003). Use of single-system research to evaluate the effectiveness of cognitive-behavioural treatment of schizophrenia. *British Journal of Social Work, 33,* 885–899.

Brainerd, C .J., &, Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press.

Brandt, A. M. (1978). Racism and research: The case of the Tuskegee Syphilis Study. *Hastings Center Report*, *8*(6), 21–29.

Braver, M. C. W., & Braver, S. L. (1988). Statistical treatment of the Solomon Four-Group Design: A meta-analytic approach. *Psychological Bulletin*, *104*, 150–154.

Brestan, E. V., & Eyberg, S. M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology*, *27*(2), 180–189.

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Boca Raton, FL: Chapman and Hall/CRC.

Breuer, J., & Freud, S. (1957). *Studies in hysteria*. New York: Basic Books.

Brim, R. L., & Miller, F. G. (2013). The potential benefit of the placebo effect in sham-controlled trials: Implications for risk-benefit assessments and informed consent. *Journal of Medical Ethics*, *39*, 703–707.

Brooks, A., Todd, A. W., Tofflemoyer, S., & Horner, R. H. (2003). Use of functional assessment and a self-management system to increase academic engagement and work completion. *Journal of Positive Behavior Interventions*, *5*, 144–152.

Brossart, D. F., Meythaler, J. M., Parker, R. I., McNamara, J, & Elliott, T. R. (2008). Advanced regression methods for single-case designs: Studying propranolol in the treatment for agitation associated with traumatic brain injury. *Journal of Rehabilitation Psychology*, *53*, 357–369.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531–563.

Brown, L. K., Hadley, W., Stewart, A., Lescano, C., Whiteley, L., Donenberg, G., & DiClemente, R. (2010). Psychiatric disorders and sexual risk among adolescents in mental health treatment. *Journal of Consulting and Clinical Psychology*, *78*(4), 590–597.

Brown, S. D. (2012). Common ground for behavioural and neuroimaging research. *Australian Journal of Psychology*, *64*(1), 4–10.

Brugge, D., & Missaghian, M. (2006). Protecting the Navajo people through tribal regulation of research. *Science and Engineering Ethics*, *12*(3), 491–507.

Bryant, A., & Charmaz, K. (2012). Grounded theory and psychological research. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 2, pp. 39–56). Washington, DC: American Psychological Association.

Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E., & Snow, R. W. (2013). Mobile phones and malaria: Modeling human and parasite travel. *Travel Medicine and Infectious Disease, 11,* 15–22.

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., . . . Ochsner, K. N. (2013). Cognitive reappraisal of emotion: A meta-analysis of human neuroimaging studies. *Cerebral Cortex*. Available at http://cercor.oxfordjournals.org/content/early/2013/06/12/cercor.bht154.short

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data. *Perspectives on Psychological Science*, *6*(1), 3–5.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*(1), 1–11.

Buros Institute. (2011). *Tests in print VIII: An index to tests, test reviews, and the literature on specific tests*. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Burton, N. W., Brown, W., & Dobson, A. (2010). Accuracy of body mass index estimated from self-reported height and weight in mid-aged Australian women. *Australian and New Zealand Journal of Public Health*, *34*(6), 620–623.

Burton, T. M. (2000, November 1). Unfavorable drug study sparks battle over publication of results. *Wall Street Journal*, Vol. CCXXXVI, No. 86, pp. B1, B4.

Busch, M. L., & Howse, R. (2009). *A (genetically modified) food fight: Canada's WTO challenge to Europe's ban on gm products*. Toronto, Ontario, Canada: Howe Institute.

Bussey, T. J, Holmes, A., Lyon, L., Mar, A. C., McAllister, K. A., Nithianantharajah, J., . . . Saksida, L. M. (2012). New translational assays for preclinical modelling of cognition in schizophrenia: The touchscreen testing method for mice and rats. *Neuropharmacology*, *62*(3), 1191–1203.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales*. Minneapolis, MN: University of Minnesota Press.

Byers, A. L., Yaffe, K., Covinsky, K. E., Friedman, M. B., & Bruce, M. L. (2010). High occurrence of mood and anxiety disorders among older adults: The National Comorbidity Survey Replication. *Archives of General Psychiatry*, *67*(5), 489–496.

Cahill, J., Barkham, M., & Stiles, W. B. (2010), Systematic review of practice-based research on psychological therapies in routine clinic settings. *British Journal of Clinical Psychology*, *49*, 421–453.

Cai, D., Pearce, K., Chen, S., & Glanzman, D. L. (2011). Protein kinase M maintains long-term sensitization and long-term facilitation in Aplysia. *Journal of Neuroscience*, *31*(17), 6421–6431.

Calandrillo, S. P. (2004). Vanishing vaccinations: Why are so many Americans opting out of vaccinating their children? *University of Michigan Journal of Law Reform*, *37*(2), 353–440.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, *3*, 1077–1078.

Calzada, E. J., Basil, S., & Fernandez, Y. (2012). What Latina mothers think of evidence-based parenting practices: A qualitative study of treatment acceptability. *Cognitive and Behavioral Practice*, *20*, 362–374.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research and teaching. In N.L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics. *Journal of the American Medical Association*, *287*(4), 473–480.

Card, N. A., & Casper, M. (2013). Meta-analysis and quantitative research synthesis. In T.D. Little (Ed.), *The Oxford handbook*

of quantitative methods (Vol. 2, pp. 701–717). New York: Oxford University Press.

Carek, P. J. Laibstain, S. E., & Carek, S. M. (2011). Exercise for the treatment of depression and anxiety. *International Journal of Psychiatry in Medicine*, *41*, 15–28

Carpenter, J. R., & Kenward, M .G. (2008). *Missing data in clinical trials—A practical guide*. Birmingham, UK: National Institute for Health Research. Available at www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1560.

Carpenter, W. T., Appelbaum, P. S., & Levine, R. J. (2003). The Declaration of Helsinki and clinical trials: A focus on placebo-controlled trials in schizophrenia. *American Journal of Psychiatry*, *160*(2), 356–362.

Carr, S. M., Lhussier, M., Forster, N., Geddes, L., Deane, K., Pennington, M., . . . Hildreth, A. (2011). An evidence synthesis of qualitative and quantitative research on component intervention techniques, effectiveness, cost-effectiveness, equity and acceptability of different versions of health-related lifestyle advisor role in improving health. NIHR Health Technology Assessment programme: Executive Summaries. Available at www.ncbi.nlm.nih.gov/pubmedhealth/PMH0014942/

Carter, G. L., Clover, K., Whyte, I. M., Dawson, A. H., & D'Este, C. (2013). Postcards from the EDge: 5-year outcomes of a randomised controlled trial for hospital-treated self-poisoning. *The British Journal of Psychiatry*, *202*(5), 372–380.

Case, L., & Smith, T. B. (2000). Ethnic representation in a sample of the literature of applied psychology. *Journal of Consulting and Clinical Psychology*, *68*, 1107–1110.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S., Harrington, H., Israel, S., . . . Moffitt, T. E. (2014). The 'p factor': One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*, 119–137.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I., . . . Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, *297*, 851–854.

Castonguay, L. G., Constantino, M. J., Boswell, J. F., & Kraus, D. R. (2011). The therapeutic alliance: Research and theory. In L.M. Horowitz & S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment and therapeutic interventions* (pp. 509–518). Hoboken, NJ: John Wiley & Sons.

Cech, T. R., & Leonard, J. S. (2001). Conflicts of interest—Moving beyond disclosure. *Science*, *291*, 989.

Centers for Disease Control and Prevention. (2009). Reduced hospitalizations for acute myocardial infarction after implementation of a smoke-free ordinance—City of Pueblo, Colorado, 2002—2006. *Morbidity and Mortality Weekly Report*, *57*(51&52); 1373–1377.

Centers for Disease Control and Prevention (2012a). Low level lead exposure harms children: A renewed call for primary prevention: Report of the Advisory Committee on Childhood Lead Poisoning Prevention. Atlanta, GA: CDC.

Centers for Disease Control and Prevention. (2012b). Prevalence of Autism Spectrum Disorders—Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. *Morbidity and Mortality Weekly Report*, *61*, 1–19.

Chambers, C. D., Garavan, H., & Bellgrove, M. A. (2009). Insights into the neural basis of response inhibition from cognitive and clinical neuroscience. *Neuroscience & Biobehavioral Reviews*, *33*(5), 631–646.

Chan, A. W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ, 330*(7494), 753.

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *Journal of the American Medical Association, 291*(20), 2457–2465.

Cheng, C., Cheung, S. F., Chio, J. H. M., & Chan, M. P. S. (2013). Cultural meaning of perceived control: A meta-analysis of locus of control and psychological symptoms across 18 cultural regions. *Psychological Bulletin*, *139*(1), 152–188.

Child Welfare Information Gateway. (2006). Long-term consequences of child abuse and neglect. Retrieved September 23, 2008, from www.childwelfare.gov/pubs/factsheets/long_term_consequences.cfm

Chiviacowsky, S., Wulf, G., Lewthwaite, R., & Campos, T. (2012). Motor learning benefits of self-controlled practice in persons with Parkinson's disease. *Gait & Posture*, *35*, 601–605.

Choi, A. N., Lee, M. S., & Lim, H. J. (2008). Effects of group music intervention on depression, anxiety, and relationships in psychiatric patients: A pilot study. *The Journal of Alternative and Complementary Medicine*, *14*(5), 567–570.

Church, D., Hawk, C., Brooks, A. J., Toukolehto, O., Wren, M., Dinter, I., & Stein, P. (2013). Psychological trauma symptom improvement in veterans using emotional freedom techniques: A randomized controlled trial. *The Journal of Nervous and Mental Disease*, *201*(2), 153–160.

Church, R. M. (1964). Systematic effect of random error in the yoked control design. *Psychological Bulletin*, *62*, 122–131.

Cicchetti, D. V. (1991). The reliability of the peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*, 119–186.

Coakley, A. B., & Mahoney, E. K. (2009). Creating a therapeutic and healing environment with a pet therapy program. *Complementary Therapies in Clinical Practice, 15,* 141–146.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American psychologist*, *45*(12), 1304–1312.

Cohen, J. (2013). More woes for struggling HIV vaccine field. *Science*, *340*(6133), 667.

Conroy, S., Marks, M. N., Schacht, R., Davies, H. A., & Moran, P. (2010). The impact of maternal depression and personality disorder on early infant care. *Social Psychiatry and Psychiatric Epidemiology*, *45*(3), 285–292.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*(1), 56–68.

Cooky, C., & Dworkin, S. L. (2013). Policing the boundaries of sex: A critical examination of gender verification and the Caster Semenya controversy. *Journal of Sex Research*, *50*(2), 103–111.

Cooper, H. (Ed.) (2012). *APA handbook of research methods in psychology*. Washington, DC: American Psychological Association.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cooper, H., & VandenBos, G. R. (2013). *Archives of Scientific Psychology*: A new journal for a new era. *Archives of Scientific Psychology*, *1*(1), 1–6.

Copeland, W. E., Wolke, D., Angold, A., & Costello, E. J. (2013). Adult psychiatric outcomes of bullying and being bullied by peers in childhood and adolescence: Psychiatric outcomes of bullying and being bullied. *JAMA Psychiatry*, *70*(4), 419–426.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 26. Available at www.ncbi.nlm.nih.gov/pmc/articles/PMC3095378/

Couzin-Frankel, J. (2012). Service offers to reproduce results for a fee. *Science*, 337, 1031.

Couzin-Frankel, J. (2013a). Complete repeat? Initiative gets $1.3 million to try to replicate cancer studies. *Science News*. Available at http://news.sciencemag.org/author/jennifer-couzin-frankel

Couzin-Frankel, J. (2013b). When mice mislead. *Science*, *342* (6161), 922–925.

Couzin-Frankel, J. (2014). Divulging DNA secrets of dead stirs debate. *Science*, *343*, 356–357.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553–558.

Coyne, J. C., Thombs, B. D., Stefanek, M., & Palmer, S. C. (2009). Time to let go of the illusion that psychotherapy extends the survival of cancer patients: Reply to Kraemer, Kuchler, and Spiegel (2009). *Psychological Bulletin*, *135*, 179–182.

Critchfield, T. S., Haley, R., Sabo, B., Colbert, J., & Macropoulis, G. (2003). A half century of scalloping in the work habits of the United States Congress. *Journal of Applied Behavior Analysis*, *36*, 465–486.

Crits-Christoph, P., Gibbons, M. B. C., & Mukherjee, D. (2013). Psychotherapy process outcome research. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 298–339). New York: John Wiley & Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, *381*(9875), 1371–1379.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley & Sons.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Taylor & Francis Group.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, *177*, 7–11.

Cunningham, T. R., & Austin, J. (2007). Using goal setting, task clarification, and feedback to increase the use of the hands-free technique by hospital operating room staff. *Journal of Applied Behavior Analysis*, *40*, 673–677.

Cuny, H. (1965). *Ivan Pavlov: The man and his theory*. (P. Evans, translation.) New York: Paul S. Eriksson Publisher.

Curcio, A. L., Mak, A. S., & George, A. M. (2012). Do adolescent delinquency and problem drinking share psychosocial risk factors? A literature review. *Addictive Behaviors*, *38*(4), 2003–2013.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 126. Available at www.biomedcentral.com/1741-7015/11/126

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology*, *32*(4), 302–314.

Damaser, E., Whitehouse, W. G., Orne, M. T., Orne, E. C., & Dinges, D. F. (2009). Behavioral persistence in carrying out a posthypnotic suggestion beyond the hypnotic context: A consideration of the role of perceived demand characteristics. *International Journal of Clinical and Experimental Hypnosis*, *58*(1), 1–20.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*, 6889–6892.

Davis, M., Myers, K. M., Chhatwal, J., & Ressler, K. J. (2006). Pharmacological treatments that facilitate extinction of fear: Relevance to psychotherapy. *NeuroRx*, *3*, 82–96.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., . . . Varley, J. (2010). Randomized controlled trial of an intervention for toddlers with Autism: The Early Start Denver Model. *Pediatrics. 125*, e17–e23.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., . . . Varley, J. (2010). Randomized controlled trial of an intervention for toddlers with Autism: The Early Start Denver Model. *Pediatrics. 125*, e17–e23.

De Angelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., . . . Weyden, M. B. V. D. (2005). Is this clinical trial fully registered?—A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *352*(23), 2436–2438.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509.

De Los Reyes, A., & Kazdin, A. E. (2006). Conceptualizing changes in behavior in intervention research: The Range of Possible Changes Model. *Psychological Review*, *113*, 554–583.

De Los Reyes, A., Kundey, S., & Wang, M. (2011). The end of the primary outcome measure: A research agenda for constructing its replacement. *Clinical Psychology Review*, *31*(5), 829–838.

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*, 123–149.

De Pisapia, N., Bornstein, M. H., Rigo, P., Esposito, G., De Falco, S., & Venuti, P. (2013). Gender differences in directional brain responses to infant hunger cries. *NeuroReport*, *24*(3), 142–146.

Deater-Deckard, K. D. (2004). *Parenting stress*. New Haven, CT: Yale University Press.

Deep-Soboslay, A., Benes, F. M., Haroutunian, V., Ellis, J. K., Kleinman, J. E., & Hyde, T. M. (2011). Psychiatric brain banking: Three perspectives on current trends and future directions. *Biological Psychiatry*, *69*, 104–112.

Deer, B. (2011). How the case against the MMR vaccine was fixed. *British Medical Journal*, *342*. *c5347*

DeMets, D. L., Furberg, C., & Friedman, L. M. (2006). *Data monitoring in clinical trials*. New York: Springer.

Denzin, N. H. & Lincoln, Y. S. (Eds.). (2011). *The Sage handbook of qualitative research* (4th ed). Thousand Oaks, CA: Sage.

Derogatis, L. R., & Unger, R. (2010). Symptom Checklist-90 Revised. In I.B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology*. Hoboken, NJ: John Wiley & Sons.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PloS One*, *9*(1), e83875.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94,* 361–366.

DeStefano, F., & Thompson, W. W. (2004). MMR vaccine and autism: An update of the scientific evidence. *Expert Review of Vaccines*, *3*(1), 19–22.

Diener, E., & Chan, M. Y. (2011). Happy people live longer: Subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being*, *3*, 1–43.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290.

DiLillo, D., Hayes-Skelton, S. A., Fortier, M. A., Perry, A. R., Evans, S. E., Messman Moore, T. L., . . . Fauchier, A. (2010). Development and initial psychometric properties of the Computer Assisted Maltreatment Inventory (CAMI): A comprehensive self-report measure of child maltreatment history. *Child Abuse & Neglect*, *34*(5), 305–317.

Doehrmann, O., Ghosh, S. S., Polli, F. E., Reynolds, G. O., Horn, F., Keshavan, A., . . . Gabrieli, J. D. (2013). Predicting treatment response in social anxiety disorder from functional magnetic resonance imaging. *JAMA Psychiatry*, *70*, 87–97.

Dominguez, D., Jawara, M., Martino, N., Sinaii, N., & Grady, C. (2012). Commonly performed procedures in clinical research: A benchmark for payment. *Contemporary Clinical Trials*, *33*(5), 860–868.

Dominus, S. (2011, April 20). The crash and burn of an autism guru. *The New York Times Magazine*. Available at www.nytimes.com/2011/04/24/magazine/mag-24Autism-t.html?pagewanted=all&_r=0

Dong, G., Lu, Q., Zhou, H., & Zhao, X. (2011). Precursor or sequela: Pathological disorders in people with internet addiction disorder. *PLoS One, 6*(2): e14703.

Donnellan, M. B., & Lucas, R. E. (2013). Secondary data analysis. In T.D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 2, pp. 665–677). New York: Oxford University Press.

Donoughe, K., Whitestone, J., & Gabler, H. C. (2012). Analysis of firetruck crashes and associated firefighter injuries in the United States. *Annals of Advances in Automotive Medicine*, *56*, 69–76.

Doohan, I., & Saveman, B. I. (2014). Impact on life after a major bus crash–a qualitative study of survivors' experiences. *Scandinavian Journal of Caring Sciences, 28*(1), 155–163.

Dorn, L. D., Sontag-Padilla, L. M., Pabst, S., Tissot, A., & Susman, E. J. (2013). Longitudinal reliability of self-reported age at menarche in adolescent girls: Variability across time and setting. *Developmental Psychology*, *49*(6), 1187–1193.

Driessen, E., Van, H. L., Don, F. J., Peen, J., Kook, S., Westra, D., . . . Dekker, J. J. M. (2013). The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: A randomized clinical trial. *American Journal of Psychiatry*, *170*(9): 1041–1050.

Druss, B. G., Hwang, I., Petukhova, M., Sampson, N. A., Wang, P. S., & Kessler, R. C. (2009). Impairment in role functioning in mental and chronic medical disorders in the US: Results from the National Comorbidity Survey Replication. *Molecular Psychiatry, 14,* 728–737.

Duman, C. H., Schlesinger, L., Russell, D. S., & Duman, R. S. (2008). Voluntary exercise produces antidepressant and anxiolytic behavioral effects in mice. *Brain Research*, *1199*, 148–158.

Duman, R. S., & Aghajanian, G. K. (2012). Synaptic dysfunction in depression: Potential therapeutic targets. *Science*, *338*, 68–72.

Duncan, B. L., Miller, S. D., Wampold, B. E., & Hubble, M. A. (Eds.). (2010). *The heart and soul of change: Delivering what works in therapy* (2nd ed.). Washington, DC: American Psychological Association.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and application* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Dunning, D., & Balcetis, E. (2013). Wishful seeing how preferences shape visual perception. *Current Directions in Psychological Science*, *22*(1), 33–37.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., . . . Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*, *3*(8), e3081.

Edelson, P. J. (2004). Henry K. Beecher and Maurice Pappworth: Honor in the development of the ethics of human experimentation. In V. Roelcke & V. Maio (Eds.), *Twentieth century ethics of human subject research—Historical perspectives on values, practices and regulations* (pp. 219–233). Stuttgart: Franz Steiner Verlag.

Editors of *The Lancet*. (2010). Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet*, *375*(9713), 445.

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S. J., & Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin*, *35*(5), 635–642.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Edwards, M., Wood, F., Davies, M., & Edwards, A. (2013). "Distributed health literacy": Longitudinal qualitative analysis of the roles of health literacy mediators and social networks of people living with a long-term health condition. *Health Expectations*. Available at http://onlinelibrary.wiley.com/doi/10.1111/hex.12093/abstract?deniedAccessCustomisedMessage=&userIsAuthenticated=false

Eggert, L. D. (2011). Best practices for allocating appropriate credit and responsibility to authors of multi-authored articles. *Frontiers in psychology*, *2*. Available at www.ncbi.nlm.nih.gov/pmc/articles/PMC3164109/

Ehrlich, P. R., & Ehrlich, A. H. (2008). *The dominant animal: Human evolution and the environment*. Washington, DC: Island Press.

Ehrlich, S., Brauns, S., Yendiki, A., Ho, B. C., Calhoun, V., Schulz, S. C., . . . Sponheim, S. R. (2012). Associations of cortical thickness and cognition in patients with schizophrenia and healthy controls. *Schizophrenia Bulletin*, *38*(5), 1050–1062.

Eichler, H. G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., . . . Bloechl-Daum, B. (2011). Bridging the efficacy–effectiveness gap: A regulator's perspective on addressing variability of drug response. *Nature Reviews Drug Discovery*, *10*(7), 495–506.

Eisner, M., Nagin, D., Ribeaud, D., & Malti, T. (2012). Effects of a universal parenting program for highly adherent parents: A propensity score matching approach. *Prevention Science*, *13*(3), 252–266.

Elkins, S. R., Moore, T. M., McNulty, J. K., Kivisto, A. J., & Handsel, V. A. (2013). Electronic diary assessment of the temporal association between proximal anger and intimate partner violence perpetration. *Psychology of Violence*, *3*(1), 100–113.

Embry, D. D. (2002). The Good Behavior Game: A best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review, 5,* 273–297.

Epp, A. M., Dobson, K. S., Dozois, D. J., & Frewen, P. A. (2012). A systematic meta-analysis of the Stroop task in depression. *Clinical Psychology Review*, *32*(4), 316–328.

Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., Kwon, T., . . . Jeong, J. S. (2011). AutoSense: Unobtrusively wearable sensor suite for inferencing of onset, causality, and consequences of stress in the field. In *Proceedings of ACM* , SenSys, Seattle, WA.

Esser, G., Schmidt, M. H., & Woerner, W. (1990). Epidemiology and course of psychiatric disorders in school-age children: Results of a longitudinal study. *Journal of Child Psychology and Psychiatry*, *31*, 243–263.

Exner, J. E., Jr., & Erdberg, P. (2005). *The Rorschach: A comprehensive system: Volume 2: Advanced interpretation*. Hoboken, NJ: John Wiley & Sons.

Falagas, M. E., Korbila, I. P., Giannopoulou, K. P., Kondilis, B. K., & Peppas, G. (2009). Informed consent: How much and what do patients understand? *The American Journal of Surgery*, *198*(3), 420–435.

Fallucca, E., MacMaster, F. P., Haddad, J., Easter, P., Dick, R., May, G., . . . Rosenberg, D. R. (2011). Distinguishing between major depressive disorder and obsessive-compulsive disorder in children by measuring regional cortical thickness. *Archives of General Psychiatry*, *68*(5), 527–533.

Fanaj, N., Poniku, I., Gashi, M., & Muja, G. (2012). P-280-Hopelessness and self-esteem of adolescents referred to mental health clinic in Prizren. *European Psychiatry*, *27*(Supplement 1), 1.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033.

Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., . . . Barlow, D. H. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A randomized controlled trial. *Behavior Therapy*, *43*(3), 666–678.

Farrington, D. P. (1991). Childhood aggression and adult violence: Early precursors and later life outcomes. In D.J. Pepler & K.H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 5–29). Hillsdale, NJ: Erlbaum.

Fast, N. J., Halevy, N., & Galinsky, A. D. (2012). The destructive nature of power without status. *Journal of Experimental Social Psychology*, *48*(1), 391–394.

Feather, J. S., & Ronan, K. R. (2006). Trauma-focused cognitive-behavioural therapy for abused children with posttraumatic stress disorder. *New Zealand Journal of Psychology*, *35*, 132–145.

Fein, D., Barton, M., Eigsti, I. M., Naigles, L., Schultz, R. T., Stevens, M., . . . Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry*, *54*, 195–205.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Ferguson, D. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, *70*, 165–178.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*(2), 119–126.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The *long way from α-error control to validity proper problems with a short-sighted false-positive debate*. *Perspectives on Psychological Science*, *7*(6), 661–669.

Field, T. (2010). Postpartum depression effects on early interactions, parenting, and safety practices: A review. *Infant Behavior and Development*, *33*(1), 1–6.

Fienberg, S. E. (Ed.). (2014). *Annual review of statistics and its application* (Vol. 1). Palo Alto, CA: Annual Reviews.

Finniss, D. G., Kaptchuk, T. J., Miller, F., & Benedetti, F. (2010). Biological, clinical, and ethical advances of placebo effects. *The Lancet*, *375*(9715), 686–695.

Fisher, L. B., Miles, I. W., Austin, S. B., Camargo Jr., C. A., & Colditz, G. A. (2007). Predictors of initiation of alcohol use among US adolescents: Findings from a prospective cohort study. *Archives of Pediatrics & Adolescent Medicine*, *161*(10), 959–966.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver & Boyd.

Flood, W. A., & Wilder, D. A. (2004). The use of differential reinforcement and fading to increase time away from a caregiver in a child with separation anxiety disorder. *Education and Treatment of Children, 27*, 1–8.

Flory, J., & Emanuel, E. (2004). Interventions to improve research participants' understanding in informed consent for research. *Journal of the American Medical Association*, *292*(13), 1593–1601.

Flückiger, C., Del Re, A. C., Wampold, B. E., Symonds, D., & Horvath, A. O. (2012). How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. *Journal of Counseling Psychology*, *59*(1), 10–17.

Forman, E. M., Shaw, J. A., Goetter, E. M., Herbert, J. D., Park, J. A., & Yuen, E. K. (2012). Long-term follow-up of a randomized controlled trial comparing acceptance and commitment therapy and standard cognitive behavior therapy for anxiety and depression. *Behavior Therapy*, *43*(4), 801–811.

Fouchier, R. A., Herfst, S., & Osterhaus, A. D. M. E. (2012). Restricted data on influenza H5N1 virus transmission. *Science*, *335*(6069), 662–663.

Foulks, E. F. (1987). Social stratification and alcohol use in North Alaska. *Journal of Community Psychology*, *15*, 349–356.

Foulks, E. F. (1989). Misalliances in the Barrow Alcohol Study. *American Indian and Native Alaska Mental Health Research*, *2* (3), 7–17.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity. *Journal of the American Medical Association*, *303*(1), 47–53.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594.

Frank, E., Cassano, G. B., Rucci, P., Thompson, W. K., Kraemer, H. C., Fagiolini, A., . . . Forgione, R. N. (2011). Predictors and moderators of time to remission of major depression with interpersonal psychotherapy and SSRI pharmacotherapy. *Psychological Medicine, 41*, 151–162.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing* (3rd ed.). Baltimore, MD: Johns Hopkins University Press.

Frank, J. D., Nash, E. H., Stone, A. R., & Imber, S. D. (1963). Immediate and long-term symptomatic course of psychiatric outpatients. *American Journal of Psychiatry*, *120*, 429–439.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.

Franke, G. H. (1999). Effects of computer administration on the Symptom Checklist (SCL-90-R) with a special focus on the item sequence. *Diagnostica*, *45*, 147–153.

Franklin, B., Jones, A., Love, D., Puckett, S., Macklin, J., & White-Means, S. (2012). Exploring mediators of food insecurity and obesity: A review of recent literature. *Journal of Community Health*, *37*(1), 253–264.

Frans, E. M., Sandin, S., Reichenberg, A., Långström, N., Lichtenstein, P., McGrath, J. J., & Hultman, C. M. (2013). Autism risk across generations: A population-based study of advancing grandpaternal and paternal ageautism risk. *JAMA Psychiatry*, *70*(5), 516–521.

Frass, M., Strassl, R. P., Friehs, H., Müllner, M., Kundi, M., & Kaye, A. D. (2012). Use and acceptance of complementary and alternative medicine among the general population and medical personnel: A systematic review. *The Ochsner Journal*, *12*(1), 45–56.

Freedland, K. E., Mohr, D. C., Davidson, K. W., & Schwartz, J. E. (2011). Usual and unusual care: Existing practice control groups in randomized controlled trials of behavioral interventions. *Psychosomatic Medicine*, *73*(4), 323–335.

Freedman, N. D., Park, Y., Abnet, C. C., Hollenbeck, A. R., & Sinha, R. (2012). Association of coffee drinking with total and cause-specific mortality. *New England Journal of Medicine*, *366*(20), 1891–1904.

Frewen, P. A., Dozois, D. J. A., & Lanius, R. A. (2008). Neuroimaging studies of psychological interventions for mood and anxiety disorders: Empirical and methodological review. *Clinical Psychology Review*, *28*, 228–246.

Frey, B. S. (2011). Happy people live longer. *Science*, *331*, 542–543.

Frisch, M. B., Cornell, J., Villanueva, M., & Retzlaff, P. J. (1992). Clinical validation of the Quality of Life Inventory. A measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological Assessment*, *4*(1), 92.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18.

Frost, N. (2011). *Qualitative research methods in psychology: From core to combined approaches*. Berkshire, England: Open University Press.

Fryers, T., & Brugha, T. (2013). Childhood determinants of adult psychiatric disorder. *Clinical Practice & Epidemiology* in *Mental Health*, *9*, 1–50.

Gamble, J. L., & Hess, J. J. (2012). Temperature and violent crime in Dallas, Texas: Relationships and implications of climate change. *Western Journal of Emergency Medicine*, *13*(3), 239–246.

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2005). Understanding the effects of task-specific practice in the brain: Insights from individual-differences analyses. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(2), 235–245.

Garcia, J. R., Reiber, C., Massey, S. G., & Merriwether, A. M. (2012). Sexual hookup culture: A review. *Review of General Psychology*, *16*(2), 161–176.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., . . . Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *Journal of the American Medical Association*, *293*(10), 1223–1238.

Geier, D. A., Kern, J. K., King, P. G., Sykes, L. K., & Geier, M. R. (2012). Hair toxic metal concentrations and autism spectrum disorder severity in young children. *International Journal of Environmental Research and Public Health*, *9*(12), 4486–4497.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*, 445–449.

Genet, J. J., & Siemer, M. (2012). Rumination moderates the effects of daily events on negative mood: Results from a diary study. *Emotion*, *12*(6), 1329–1339.

Gentili, C., Ricciardi, E., Gobbini, M. I., Santarelli, M. F., Haxby, J. V., Pietrini, P., & Guazzelli, M. (2009). Beyond amygdala: Default mode network activity differs between patients with social phobia and healthy controls. *Brain Research Bulletin*, *79*(6), 409–413.

Gervais, W. M., & Norenzayan, A. (2012). Like a camera in the sky? Thinking about God increases public self-awareness and socially desirable responding. *Journal of Experimental Social Psychology*, *48*(1), 298–302.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*(4), 361–368.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104–1112.

Gladis, M. M., Gosch, E. A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 320–331.

Gladwell, M. (2008). *Outliers: The story of success*. New York: Little, Brown, & Company.

Glaser, R., & Kiecolt-Glaser, J. K. (2005). Stress-induced immune dysfunction: Implications for health. *Nature Reviews Immunology*, *5*(3), 243–251.

Glassman, A. H., Bigger Jr, J. T., & Gaffney, M. (2009). Psychiatric characteristics associated with long-term mortality among 361 patients having an acute coronary syndrome and major depression: Seven-year follow-up of SADHART participants. *Archives of General Psychiatry*, *66*(9), 1022–1029.

Glenn, B. A., Bastani, R., & Maxwell, A. E. (2013). The perils of ignoring design effects in experimental studies: Lessons from a mammography screening trial. *Psychology & Health*, *28*(5), 593–602.

Glenn, I. M., & Dallery, J. (2007). Effects of internet-based voucher reinforcement and a transdermal nicotine patch on cigarette smoking. *Journal of Applied Behavior Analysis, 40,* 1–13.

Goodman, S. H., Rouse, M. H., Connell, A. M., Broth, M. R., Hall, C. M., & Heyward, D. (2011). Maternal depression and child psychopathology: A meta-analytic review. *Clinical Child and Family Psychology Review*, *14*(1), 1–27.

Goodman, S. R., & Mallet, R. T. (2012). Plagiarism. *Experimental Biology and Medicine*, *237*(7), 739.

Gorgolewski, K. J., Margulies, D. S., & Milham, M. P. (2013). Making data sharing count: A publication-based solution. *Frontiers in Neuroscience*, *7*. Available at www.ncbi.nlm.nih.gov/pmc/articles/PMC3565154/#__ffn_sectitle

Gottschalk, L. A. & Bechtel, R. J. (Eds.). (2009). *Computerized content analysis of speech and verbal texts and its many applications.* Hauppauge, NY: Nova Science Publishers.

Gottschalk, L. A., Bechtel, R. J., Maguire, G. A., Harrington, D. E., Levinson, D. M., Franklin, D. L., & Carcamo, D. (2000). Computerized measurement of cognitive impairment and associated neuropsychiatric dimensions. *Comprehensive Psychiatry*, *41*, 326–333.

Grady, C. (2012). Undue worry about paying research participants? *Clinical Investigation*, *2*(9), 855–857.

Graham, S. (1992). "Most of the subjects were White and middle class:" Trends in published research on African Americans in selected APA journals 1970–1989. *American Psychologist*, *47*, 629–639.

Grant, D. A. (1948). The Latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin*, *45*, 427–442.

Gray, F. D. (1998). *The Tuskegee Syphilis Study: The real story and beyond*. Montgomery, AL: Newsouth Books.

Grecco, E., Robbins, S. J., Bartoli, E., & Wolff, E. F. (2013). Use of nonconscious priming to promote self-disclosure. *Clinical Psychological Science*, *1*(3), 311–315.

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, *53*(4–5), 225–228.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99–108.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Grimm, D. (2010). Is a dolphin a person? *Science,* 327, 1070–1071.

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Volume 1: Foundations, planning, measures, and psychometrics* (pp. 621–642). Washington, DC: American Psychological Association.

Grissom, R. J. (1996). The magical number .7 +/− .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology*, *64*, 973–982.

Grissom, R. J., & Kim. J. J. (2011). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Taylor & Francis.

Gross, L. (2009). A broken trust: Lessons from the vaccine–autism wars. *PLoS Biology*, *7*(5), e1000114.

Grossman, A. H., Haney, A. P., Edwards, P., Alessi, E. J., Ardon, M., & Howell, T. J. (2009). Lesbian, gay, bisexual and transgender youth talk about experiencing and coping with school violence: A qualitative study. *Journal of LGBT Youth*, *6*(1), 24–46.

Groth-Marnat, G. (Ed.) (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Grundy, C. T., Lunnen, K. M., Lambert, M. J., Ashton, J. E., & Tovey, D. R. (1994). The Hamilton Rating Scale for Depression: One scale or many? *Clinical Psychology: Science and Practice*, *1*, 197–205.

Guest, G. (2013). Describing mixed methods research: An alternative to typologies. *Journal of Mixed Methods Research*, *7*, 141–151.

Guilford, J. M., & Pezzuto, J. M. (2011). Wine and health: A review. *American Journal of Enology and Viticulture*, *62*, 471–486.

Gullickson, A., & Morning, A. (2011). Choosing race: Multiracial ancestry and identification. *Social Science Research*, *40*(2), 498–512.

Gunter, W. D., & Daly, K. (2012). Causal or spurious: Using propensity score matching to detangle the relationship between violent video games and violent behavior. *Computers in Human Behavior*, *28*(4), 1348–1355.

Gunther, A. (2011). PSYCLINE: Your guide to psychology and social science journals on the Web. Retrieved August 2011 from www.psycline.org/journals/psycline.html

Guthrie, R. V. (2003). *Even the rat was white: A historical view of psychology* (2nd ed). Upper Saddle River, NJ: Pearson Education.

Guxens, M., Aguilera, I., Ballester, F., Estarlich, M., Fernández-Somoano, A., Lertxundi, A., . . . Sunyer, J. (2012). Prenatal exposure to residential air pollution and infant mental development: Modulation by antioxidants and detoxification factors. *Environmental Health Perspectives*, *120*(1), 144–149.

Guyatt, G., Akl, E. A., Hirsh, J., Kearon, C., Crowther, M., Gutterman, D., . . . Schnemann, H. (2010). The vexing problem of guidelines and conflict of interest: A potential solution. *Annals of Internal Medicine*, *152*(11), 738–741.

Haas, S. A., Krueger, P. M., & Rohlfsen, L. (2012). Race/ethnic and nativity disparities in later life physical performance: The role of health and socioeconomic status over the life course. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *67*(2), 238–248.

Haller, G., Haller, D. M., Courvoisier, D. S., & Lovis, C. (2009). Handheld vs. laptop computers for electronic data collection in clinical research: A crossover randomized trial. *Journal of the American Medical Informatics Association*, *16*(5), 651–659.

Hallfors, D., Khatapoush, S., Kadushin, C., Watson, K., & Saxe, L. (2000). A comparison of paper vs. computer-assisted self-interview for school alcohol, tobacco, and other drug surveys. *Evaluation and Program Planning*, *23*, 149–155.

Hamilton, J. P., Etkin, A., Furman, D. J., Lemus, M. G., Johnson, R. F., & Gotlib, I. H. (2012). Functional neuroimaging of major depressive disorder: A meta-analysis and new integration of baseline activation and neural response data. *American Journal of Psychiatry*, *169*(7), 693–703.

Hardcastle, S. J., Taylor, A. H., Bailey, M. P., Harley, R. A., & Hagger, M. S. (2013). Effectiveness of a motivational interviewing intervention on weight loss, physical activity and cardiovascular disease risk factors: A randomised controlled trial with a 12-month post-intervention follow-up. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 40. Available at www.ijbnpa.org/content/10/1/40/

Harding, A., Harper, B., Stone, D., O'Neill, C., Berger, P., Harris, S., & Donatuto, J. (2012). Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. *Environmental Health Perspectives*, *120*(1), 6–10.

Hardt, J., & Rutter, M. (2004). Validity of adult retrospective reports of adverse childhood experiences: Review of the evidence. *Journal of Child Psychology and Psychiatry*, *45*(2), 260–273.

Harkness, E., Macdonald, W., Valderas, J., Coventry, P., Gask, L., & Bower, P. (2010). Identifying psychosocial interventions that improve both physical and mental health in patients with diabetes: A systematic review and meta-analysis. *Diabetes Care*, *33*(4), 926–930.

Harkness, J., Lederer, S. E., & Wikler, D. (2001). Laying ethical foundations for clinical research. *Bulletin of the World Health Organization*, *79*(4), 365–366.

Harley, T. A. (2004). Does cognitive neuropsychology have a future? *Cognitive Neuropsychology*, *21*(1), 3–16.

Harwood, T. M., & L'Abate, L. (2010). *Self-help in mental health: A critical review.* New York: Springer.

Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.) (2005). *The new unconscious*. New York: Guilford Press.

Hawkley, L. C., & Cacioppo, J. T. (2010). Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of Behavioral Medicine*, *40*(2), 218–227.

Hawton, K., Casañas i Comabella, C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: A systematic review. *Journal of Affective Disorders*, *147*, 17–28.

Hedden, T., Ketay, S., Aron, A., Markus, H. R., & Gabrieli, J. D. (2008). Cultural influences on neural substrates of attentional control. *Psychological science*, *19*(1), 12–17.

Heilbronn, L. K., & Ravussin, E. (2003). Calorie restriction and aging: Review of the literature and implications for studies in humans. *American Journal of Clinical Nutrition*, *78*(3), 361–369.

Heinz, A. J., Beck, A., Meyer-Lindenberg, A., Sterzer, P., & Heinz, A. (2011). Cognitive and neurobiological mechanisms of alcohol-related aggression. *Nature Reviews Neuroscience*, *12*(7), 400–413.

Henley, N. M. (1977). *Body politics: Power, sex, and nonverbal communication*. Englewood Cliffs, NJ: Prentice-Hall.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Herson, J., Buyse, M., & Wittes, J. T. (2012). On stopping a randomized clinical trial for futility. In D. Harrington (Ed.), *Designs for clinical trials* (pp. 109–137). New York: Springer Science + Business Media.

Hertenstein, M. J., & Weiss, S. J. (Eds.). (2011). *The handbook of touch: Neuroscience, behavioral, and health perspectives*. New York: Springer.

Hibbard, S. (2003). A critique of Lilienfeld et al.'s (2000). The scientific status of projective techniques. *Journal of Personality Assessment*, *80*(3), 260–271.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300.

Hill, A. K., Hunt, J., Welling, L. L., Cárdenas, R. A., Rotella, M. A., Wheatley, J. R., . . . Puts, D. A. (2013). Quantifying the strength and form of sexual selection on men's traits. *Evolution and Human Behavior*, *34*(5), 334–341.

Hill, C. E., Chui, H., & Baumann, E. (2013). Revisiting and reenvisioning the outcome problem in psychotherapy: An argument to include individualized and qualitative measurement. *Psychotherapy*, *50*, 68–76.

Hill, C. E., Chui, H., Huang, T., Jackson, J., Liu, J., & Spangler, P. (2011). Hitting the wall: A case study of interpersonal changes in psychotherapy. *Counselling and Psychotherapy Research*, *11*(1), 34–42.

Hinton, D. E., Chhean, D., Pich, V., Safren, S. A., Hofmann, S. G., & Pollack, M. H. (2005). A randomized controlled trial of cognitive-behavior therapy for Cambodian refugees with treatment-resistant PTSD and panic attacks: A cross-over design. *Journal of Traumatic Stress*, *18*(6), 617–629.

Hirsch, J. K., Visser, P. L., Chang, E. C., & Jeglic, E. L. (2012). Race and ethnic differences in hope and hopelessness as moderators of the association between depressive symptoms and suicidal behavior. *Journal of American College Health*, *60*(2), 115–125.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, *4*, 45–46.

Ho-Chunk Nation. (2008). Ho-Chunk Nation Institutional Review Board. Retrieved October 2013 from www.ho-chunknation.com/?PageId=1049

Höfler, M. (2005). The Bradford Hill considerations on causality: A counterfactual perspective. *Emerging Themes in Epidemiology*, 2, 11. Available at www.ete-online.com/content/2/1/11

Hofmann, S. G., Sawyer, A. T., Witt, A. A., & Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of Consulting and Clinical psychology*, *78*(2), 169–183.

Högberg, L., Lundholm, C., Cnattingius, S., Öberg, S., & Iliadou, A. N. (2013). Birthweight discordant female twins and their offspring: Is the intergenerational influence on birthweight due to genes or environment? *Human Reproduction*, *28*(2), 480–487.

Holland, S. (Ed.). (2012). *Arguing about bioethics*. New York City: Routledge.

Holmes, W. S. (2014). *Using propensity scores in quasi-experimental design*. Thousand Oaks, CA: Sage.

Honekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 199–209.

Hooley, J. M. (2007). Expressed emotion and relapse of psychopathology. *Annual Review of Clinical Psychology, 3*, 329–352.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews,* Issue 2. Art No: MR000010.

Horvath, A. O., & Bedi, R. P. (2002). The alliance. In J.C. Norcross (Ed.), *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* (pp. 37–69). New York: Oxford University Press.

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*, 9–16.

Howard, J. S., Mattacola, C. G., Howell, D., & Latterman, C. (2011). Response shift theory: An application for health-related quality of life in rehabilitation research and practice. *Journal of Allied Health*, *40*(1), 31–38.

Hox, J., & Balluerka, N. (2009). Special Issue: The multitrait-multimethod matrix at 50! *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*(3), 71–111.

Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *57*, 131–137.

Huey, S. J., Jr., Henggeler, S. W., Brondino, M. J., & Pickrel, S. G. (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology*, *68*, 451–467.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2007). *Designing clinical research* (3rd ed.). Philadelphia, PA: Lippincott, Williams, and Wilkins.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, *9*, 61–89.

Insel, T. (2013, November 22). Culture clash. *Director's Blog*. National Institute of Mental Health. Available at www.nimh.nih.gov/about/director/2013/culture-clash.shtml?utm_source=govdelivery&utm_medium=email&utm_campaign=govdelivery

International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine*, *126*, 36–47. Available at www.acponline.org/journal/annals/01jan97/unlfreqr.htm

International Committee of Medical Journal Editors. (2013). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: Roles and responsibilities of authors, contributors, reviewers, editors, publishers, and owners: Defining the role of authors and contributors. www.icmje.org/roles_a.html

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Iyer, G., Hanrahan, A. J., Milowsky, M. I., Al-Ahmadie, H., Scott, S. N., Janakiraman, M., . . . Solit, D. B. (2012). Genome sequencing identifies a basis for everolimus sensitivity. *Science*, *338*(6104), 221.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). New York: John Wiley & Sons.

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy: How well can clinical trials do the job. *American Psychologist*, *51*, 1031–1039.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. (1999). Methods for defining and determining the clinical significance of treatment effects in mental health research: Current status, new applications, and future directions. *Journal of Consulting and Clinical Psychology*, *67*, 300–307.

Jakubowski, M. (2011). Low-level environmental lead exposure and intellectual impairment in children—The current concepts of risk assessment. *International Journal of Occupational Medicine and Environmental Health*, *24*(1), 1–7.

Jang, Y., Kwag, K. H., & Chiriboga, D. A. (2010). Not saying I am happy does not mean I am not: Cultural influences on responses to positive affect items in the CES-D. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *65*(6), 684–690.

Janik, V. M. (2000). Whistle matching in wild bottlenose dolphins (Tursiops truncatus). *Science*, *289*, 1355–1357.

Jensen, P. S., Watanabe, H. K., & Richters, J. E. (1999). Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology*, *27*, 439–445.

Jiang, H., Emmerton, L., & McKauge, L. (2013). Academic integrity and plagiarism: A review of the influences and risk situations for health students. *Higher Education Research & Development*, *32*(3), 369–380.

Jimerson, S. R., Swearer, S. M., & Espelage, D. L. (Eds.). (2009). *Handbook of bullying in schools: An international perspective*. New York: Routledge.

Jitlal, M., Khan, I., Lee, S. M., & Hackshaw, A. (2012). Stopping clinical trials early for futility: Retrospective analysis of several randomised clinical studies. *British Journal of Cancer*, *107*(6), 910–917.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Johnson, D. O. (2013). Ethical consideration of human research. *Novel Science International Journal of Medical Sciences*, *1*(11–12).

Johnson, M. K. (2006). Memory and reality. *American Psychologist*, *61*, 760–771.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133.

Johnson, R. T., & Dickersin, K. (2007). Publication bias against negative results from clinical trials: Three of the seven deadly sins. *Nature Clinical Practice Neurology*, *3*(11), 590–591.

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience*, *37*(3), 149–152.

Jorm, A. F. (2012). Mental health literacy: Empowering the community to take action for better mental health. *American Psychologist*, *67*, 231–243.

Jose, P. E. (2008). Moderation/Mediation Help Centre. Victoria University of Wellington. Retrieved January 2013 from www.victoria.ac.nz/psyc/paul-jose-files/helpcentre/help3_mediation_background.php

Jouriles, E. N., Simpson Rowe, L., McDonald, R., Platt, C. G., & Gomez, G. S. (2011). Assessing women's responses to sexual threat: Validity of a virtual role-play Procedure. *Behavior Therapy*, *42*(3), 475–484.

Kahn, J. O., Cherng, D. W., Mayer, K., Murray, H., & Lagakos, S. (2000). Evaluation of HIV-1 immunogen, an immunologic modifier, administered to patients infected with HIV having 300 to 549 106/L CD4 cell counts: A randomized controlled trial. *Journal of the American Medical Association*, *284*, 2193–2202.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kaiser, J. (2013a). Agency nixes deCODE's new data-mining plan. *Science*, *340*, 1388–1399.

Kaiser, J. (2013b). Rare cancer successes spawn "exceptional" research efforts. *Science, 340,* 263.

Kan, P., Simonsen, S. E., Lyon, J. L., & Kestle, J. R. (2008). Cellular phone use and brain tumor: A meta-analysis. *Journal of Neuro-oncology*, *86*(1), 71–78.

Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues* (8th ed.). Belmont, CA: Wadsworth, Cengage Learning.

Kaplan, S. L. (2011). *Your child does not have Bipolar Disorder*. Santa Barbara, CA: Praeger.

Kaptchuk, T. J., & Phillips, R. S. (2011). Use of complementary and alternative medicine and self-rated health status: Results from a national survey. *Journal of General Internal Medicine*, *26*(4), 399–404.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709–1716.

Kawachi, I., Adler, N. E., & Dow, W. H. (2010). Money, schooling, and health: Mechanisms and causal evidence. *Annals of the New York Academy of Sciences*, *1186*, 56–68.

Kazantzis, N. (2000). Power to detect homework effects in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *68*, 166–170.

Kazdin, A. E. (1978). *History of behavior modification*. Baltimore, MD: University Park Press.

Kazdin, A. E. (2000). *Psychotherapy for children and adolescents: Directions for research and practice.* New York: Oxford University Press.

Kazdin, A. E. (2005). *Parent management training: Treatment for oppositional, aggressive, and antisocial behavior in children and adolescents.* New York: Oxford Press.

Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist*, *61*, 42–49.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*, 1–27.

Kazdin, A. E. (2008a). Evidence-based treatments and delivery of psychological services: Shifting our emphasis to increase impact. *Psychological Services*, *5*, 201–215.

Kazdin, A. E. (2008b). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*, 146–159.

Kazdin, A. E. (2010). Problem-solving skills training and parent management training for oppositional defiant disorder and conduct disorder. In J.R. Weisz & A.E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (2nd ed., pp. 211–226). New York: Guilford Press.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kazdin, A. E. (2013a). *Behavior modification in applied settings* (7th ed.). Long Grove, IL: Waveland Press.

Kazdin, A. E. (2013b). Evidence-based treatment and usual care: Cautions and qualifications. *JAMA Psychiatry, 70,* 666–667.

Kazdin, A. E. (2014). Moderators, mediators, and mechanisms of change in psychotherapy. In W. Lutz & S. Knox (Eds.), *Quantitative and qualitative methods in psychotherapy* (pp. 87–101). East Sussex, UK: Routledge.

Kazdin, A.E. (2015). Psychosocial treatments for conduct disorder in children and adolescents. In P.E. Nathan & J.M. Gorman (Eds.), *A guide to treatments that work* (4th ed., pp. 141–173). New York: Oxford University Press.

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *57*, 138–147.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*, 21–37.

Kazdin, A. E., French, N. H., Unis, A. S., Esveldt-Dawson, K., & Sherick, R. B. (1983). Hopelessness, depression and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Consulting and Clinical Psychology*, *51*, 504–510.

Kazdin, A. E., Holland, L., & Crowley, M. (1997). Family experience of barriers to treatment and premature termination from child therapy. *Journal of Consulting and Clinical Psychology*, *65*, 453–463.

Kazdin, A. E., Holland, L., Crowley, M., & Breton, S. (1997). Barriers to Participation in Treatment Scale: Evaluation and validation in the context of child outpatient treatment. *Journal of Child Psychology and Psychiatry*, *38*, 1051–1062.

Kazdin, A. E., & Nock, M. K. (2003). Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. *Journal of Child Psychology and Psychiatry, 44,* 1116–1129.

Kazdin, A. E., & Rabbitt, S. (2013). Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science, 1,* 170–191.

Kazdin, A. E., Rodgers, A., & Colbus, D. (1986). The Hopelessness Scale for Children: Psychometric characteristics and concurrent validity. *Journal of Consulting and Clinical Psychology*, *54*, 241–245.

Kazdin, A. E., & Rotella, C. (2008). *The Kazdin Method for parenting the defiant child: With no pills, no therapy, no contest of wills*. Boston: Houghton Mifflin.

Kazdin, A. E., & Rotella, C. (2013). *The everyday parenting toolkit: The Kazdin Method for easy, step-by-step lasting change for you and your child.* Boston: Houghton Mifflin Harcourt.

Kazdin, A. E., & Wassell, G. (2000). Therapeutic changes in children, parents, and families resulting from treatment of children with conduct problems. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 414–420.

Kazdin, A. E., & Whitley, M. K. (2003). Treatment of parental stress to enhance therapeutic change among children referred for aggressive and antisocial behavior. *Journal of Consulting and Clinical Psychology*, *71*, 504–515.

Kazdin, A. E., & Whitley, M. K. (2006). Comorbidity, case complexity, and effects of evidence-based treatment for children referred for disruptive behavior. *Journal of Consulting and Clinical Psychology, 74,* 455–467.

Kellam, S. G., Reid, J., & Balster, R. L. (2008). Effects of a universal classroom behavior program to first and second grades on young adult problem outcomes. *Drug and Alcohol Dependence, 95S,* S1–S4.

Kellehear, A. (1993). *The unobtrusive researcher: A guide to methods*. St. Leonards, Australia: Allen and Unwin.

Kempf, M. C., McLeod, J., Boehme, A. K., Walcott, M. W., Wright, L., Seal, P. . . . Moneyham, L. (2010). A qualitative study of the barriers and facilitators to retention-in-care among HIV-positive women in the rural southeastern United States: Implications for targeted interventions. *AIDS patient care and STDs*, *24*(8), 515–520.

Kendler, K. S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Archives of General Psychiatry*, *60*(9), 929–937.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-analytic Reporting Standards). *Journal of Business and Psychology*, *28*(2), 123–143.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., . . . Wang, P. S. (2009). The global burden of mental disorders: An update from the WHO World Mental Health (WMH) Surveys. *Epidemiologia e Psichiatria Sociale*, *18*, 23–33.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., . . . Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research*, *13*(2), 69–92.

Kessler, R. C., & Wang, P. S. (2008). The descriptive epidemiology of commonly occurring mental disorders in the United States. *Annual Review of Public Health, 29,* 115–129.

Kew, O. M., Sutter, R. W., de Gourville, E. M., Dowdle, W. R., & Pallansch, M. A. (2005). Vaccine-derived polioviruses and the endgame strategy for global polio eradication. *Annual Review of Microbiology*, *59*, 587–635.

Khurana, V. G., Teo, C., Kundi, M., Hardell, L., & Carlberg, M. (2009). Cell phones and brain tumors: A review including the long-term epidemiologic data. *Surgical Neurology*, *72*(3), 205–214.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*, 377–380.

Kim, Y. S., Leventhal, B. L., Koh, Y. J., Fombonne, E., Laska, E., Lim, E. C., . . . Grinker, R. R. (2011). Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry*, *168*(9), 904–912.

Kimenju, S. C., & De Groote, H. (2008). Consumer willingness to pay for genetically modified food in Kenya. *Agricultural Economics*, *38*(1), 35–46.

Kirk, R. E. (1994). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Wadsworth.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Kirpinar, I., & Oral, M. (2012). Global assessment of functioning and associated factors in psychiatric inpatients: A retrospective study. *Anadolu Psikiyatri Dergisi-Anatolian Journal of Psychiatry*, *13*(3), 198–204.

Kitayama, S., & Park, J. (2010). Cultural neuroscience of the self: Understanding the social grounding of the brain. *Social Cognitive and Affective Neuroscience*, *5*(2–3), 111–129.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Klonsky, D. E., Kotov, R., Bakst, S., Rabinowitz, J., & Bromet, E. J. (2012). Hopelessness as a predictor of attempted suicide among first admission patients with psychosis: A 10-year cohort study. *Suicide and Life-Threatening Behavior*, *42*(1), 1–10.

Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms. *Psychological Assessment*, *12*, 245–252.

Koegel, R. L. & Kern-Koegel, L. (2006). *Pivotal response treatments for autism: Communication, social, and academic development.* Baltimore: Paul H. Brookes.

Kolb, J., & Kolb, J. (2013). *The big data revolution. The world is changing. Are you ready?* Plainfield, IL: Applied Data Labs.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614.

Kosciw, J. G., Greytak, E. A., Bartkiewicz, M. J., Boesen, M. J., & Palmer, N. A. (2012). *The 2011 National School Climate Survey: The experiences of lesbian, gay, bisexual and transgender youth in our nation's schools*. New York: Gay, Lesbian and Straight Education Network (GLSEN). Report available at http://files.eric.ed.gov/fulltext/ED535177.pdf

Koslow, S. H. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience*, *3*(9), 863–865.

Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: A parametric approach. *Statistics in Medicine*, New York: John Wiley & Sons. Available at http://onlinelibrary.wiley.com/doi/10.1002/sim.5734/full

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: clinical, research, and policy importance. *Journal of the American Medical Association*, *296*(10), 1286–1289.

Kraemer, W. (2010, November). The cult of statistical significance (November 18, 2010). CESifo Working Paper Series No. 3246. Available at SSRN: http://ssrn.com/abstract=1711108

Kragh, H. (2013). The most philosophically important of all the sciences: Karl Popper and physical cosmology. *Perspectives on Science*, *21*, 325–357.

Kramer, A. D. (2010, April). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 287–290). New York: Association for Computing Machinery.

Kramer, B. S., Berg, C. D., Aberle, D. R., & Prorok, P. C. (2011). Lung cancer screening with low-dose helical CT: Results from the National Lung Screening Trial (NLST). *Journal of Medical Screening*, *18*(3), 109–111.

Kratochwill, T. R., & Shernoff, E. S. (2004). Evidence-based practice: Promoting evidence-based interventions in school psychology. *School Psychology Review*, *33*(1), 34–48.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297–300.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*, 16–26.

Krug, E. G., Dahlberg, L. L., Mercy, J. A., Zwi, A. B., & Lozano, R. (2002). *World report on violence and health*. Geneva, Switzerland: World Health Organization.

Kruschke, J. K. (2011a). *Doing Bayesian data analysis: A tutorial with R and BUGS now with JAGS*! New York: Academic Press.

Kruschke, J. K. (2011b). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science*, *6*, 272–273. (Followed by a series of articles on the topic.)

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kumar, S., Nilsen, W., Pavel, M., & Srivastava, M. (2013, January). Mobile health: Revolutionizing healthcare through transdisciplinary research. *IEEEComputer*, *46*(1): 28–35.

Kume, S., Uzu, T., Horiike, K., Chin-Kanasaki, M., Isshiki, K., Araki, S. I., . . . Koya, D. (2010). Calorie restriction enhances cell adaptation to hypoxia through Sirt1-dependent mitochondrial autophagy in mouse aged kidney. *Journal of Clinical Investigation*, *120*(4), 1043–1055.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635–642.

Kutner, B., Wilkins, C., & Yarrow, P. R. (1952). Verbal attitudes and overt behavior involving racial prejudice. *Journal of Abnormal and Social Psychology*, *47*, 649–652.

La Greca, A. M., Silverman, W. K., Lai, B., & Jaccard, J. (2010). Hurricane-related exposure experiences and stressors, other life events, and social support: Concurrent and prospective impact on children's persistent posttraumatic stress symptoms. *Journal of Consulting and Clinical Psychology*, *78*(6), 794–805.

La Greca, A. M., Silverman, W. K., Vernberg, E. M., & Prinstein, M. J. (1996). Symptoms of posttraumatic stress in children after Hurricane Andrew: A prospective study. *Journal of Consulting and Clinical Psychology*, *64*, 712–723.

La Piere, R. T. (1934). Attitudes vs. action. *Social Forces*, *13*, 230–237.

Lac, A., & Crano, W. D. (2009). Monitoring matters meta-analytic review reveals the reliable linkage of parental monitoring with adolescent marijuana use. *Perspectives on Psychological Science*, *4*(6), 578–586.

LaChance, H., Feldstein Ewing, S. W., Bryan, A. D., & Hutchison, K. E. (2009). What makes group MET work? A randomized controlled trial of college student drinkers in mandated alcohol diversion. *Psychology of Addictive Behaviors*, *23*(4), 598–612.

Laessoe, U., Hoeck, H. C., Simonsen, O., Sinkjaer, T., & Voigt, M. (2007). Fall risk in an active elderly population—can it be assessed? *Journal of Negative Results in Biomedicine*, *6*, 2.

Lai, B. S., La Greca, A. M., Auslander, B. A., & Short, M. B. (2012). Children's symptoms of posttraumatic stress and depression after a natural disaster: Comorbidity and risk factors. *Journal of Affective Disorders*, *146*, 71–78.

Laine, C., Goodman, S. N., Griswold, M. E., & Sox, H. C. (2007). Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine*, *146*(6), 450–453.

Laine, C., Horton, R., DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Godlee, F., . . . Verheugt, F. W. (2007). Clinical trial registration—Looking back and moving ahead. *New England Journal of Medicine*, *356*(26), 2734–2736.

Lall, V. F., & Levin, J. R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests. *Journal of School Psychology*, *42*, 61–86.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Client-focused research: Using client outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172.

Lambert, M. J., & Ogles, B. M. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 169–218). New York: John Wiley & Sons.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy*, *48*(1), 72–79.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10,* 288–301.

Lane, F. C., To, Y. M., Shelley, K., & Henson, R. K. (2013). An illustrative example of propensity matching with education research. *Career and Technical Education Research*, *37*, 187–212.

Langan, M. (2011). Parental voices and controversies in autism. *Disability & Society*, *26*(2), 193–205.

Lavallee, L. F. (2009). Practical application of an Indigenous research framework and two qualitative Indigenous research methods: Sharing circles and Anishnaabe symbol-based reflection. *International Journal of Qualitative Methods*, *8*(1), 21–40.

Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., . . . Das, R. (2014). RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2122–2127.

Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D. S., . . . Deisseroth, K. (2010). Global and local fMRI signals driven by neurons defined optogenetically by type and wiring. *Nature*, *465*(7299), 788–792.

Leech, N. L., & Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *Journal of Counseling & Development*, *88*(1), 61–69.

Leentjens, A. F., & Levenson, J. L. (2013). Ethical issues concerning the recruitment of university students as research subjects. *Journal of Psychosomatic Research*. 75, 394–398.

Lehrer, J. (2010). The truth wears off. *The New Yorker*. Available at www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Lehti, V., Hinkka-Yli-Salomäki, S., Cheslack-Postava, K., Gissler, M., Brown, A. S., & Sourander, A. (2013). The risk of childhood autism among second-generation migrants in Finland: A case–control study. *BMC Pediatrics*, *13*(1), 171. Available at www.biomedcentral.com/1471-2431/13/171

Lehto, S. M., Tolmunen, T., Joensuu, M., Saarinen, P. I., Valkonen-Korhonen, M., Vanninen, R., . . . Jehtonen, J. (2008). Changes in midbrain serotonin transporter availability in atypically depressed subjects after one year of psychotherapy. *Progress in Neuropsychopharmacology and Biological Psychiatry*, *32*, 229–237.

Leichsenring, F., Rabung, S., & Leibing, E. (2004). The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: A meta-analysis. *Archives of General Psychiatry*, *61*(12), 1208–1216.

Leichsenring, F., Salzer, S., J. Hilsenroth, M., Leibing, E., Leweke, F., & Rabung, S. (2011). Treatment integrity: An unresolved issue in psychotherapy research. *Current Psychiatry Reviews*, *7*(4), 313–321.

Leong, F. T. L., & Kalibatseva, Z. (2013). Clinical research with culturally diverse populations. In J.S. Comer & P.C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 413–433). New York: Oxford University Press.

Leong, F. T. L., Schmitt, N., & Lyons, B. J. (2012). Developing testable and important research questions. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 119–132). Washington, DC: American Psychological Association.

Lerman, R., BGS, H. R., Gellish, R., & Vicini, F. (2012). Improving symptoms and quality of life of female cancer survivors: A randomized controlled study. *Annals of Surgical Oncology*, *19*(2), 373–378.

Lesser, I. M., Myers, H. F., Lin, K. M., Bingham Mira, C., Joseph, N. T., Olmos, N. T., . . . Poland, R. E. (2010). Ethnic differences in antidepressant response: A prospective multi-site clinical trial. *Depression and Anxiety*, *27*(1), 56–62.

Lesser, L. I., Ebbeling, C. B., Goozner, M., Wypij, D., & Ludwig, D. S. (2007). Relationship between funding source and conclusion among nutrition-related scientific articles. *PLoS Medicine*, *4*(1), e5.

Levelt, Noort, and Drenth Committees (2012, November). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Available at www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2012/11/120695_Rapp_nov_2012_UK_web.pdf

Levesque, M., Savard, J., Simard, S., Gauthier, J. G., & Ivers, H. (2004). Efficacy of cognitive therapy for depression among women with metastatic cancer: A single-case experimental study. *Journal of Behavior Therapy and Experimental Psychiatry*, *35*, 287–305.

Lewis-Fernández, R., & Díaz, N. (2002). The cultural formulation: A method for assessing cultural factors affecting the clinical encounter. *Psychiatric Quarterly*, *73*(4), 271–295.

LeWitt, P. A., Rezai, A. R., Leehey, M. A., Ojemann, S. G., Flaherty, A. W., Eskandar, E. N., . . . Feigin, A. (2011). AAV2-*GAD* gene therapy for advanced Parkinson's disease: A double-blind, sham-surgery controlled, randomised trial. *The Lancet Neurology*, *10*(4), 309–319.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.

Lichtman, J. H., Bigger, J. T., Blumenthal, J. A., Frasure-Smith, N., Kaufmann, P. G., Lespérance, F., . . . Froelicher, E. S. (2008). Depression and coronary heart disease recommendations for screening, referral, and treatment: a science advisory from the American Heart Association Prevention Committee of the Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Epidemiology and Prevention, and Interdisciplinary Council on Quality of Care and Outcomes Research: Endorsed by the American Psychiatric Association. *Circulation*, *118*(17), 1768–1775.

Lilienfeld, S., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

Lin, K., & Poland, R. E. (2000). *Ethnicity, culture, and psychopharmacology*. American College of Neuropharmacology, Available at www.acnp.org/g4/GN401000184/CH180.html

Lind, J. (1753). *A treatise on the scurvy*. London: A. Millar.

Linden, D. E. J. (2006). How psychotherapy changes the brain—The contribution of functional neuroimaging. *Molecular Psychiatry*, *11*, 528–538.

Lindner, P., Andersson, G., Öst, L. G., & Carlbring, P. (2013). Validation of the Internet-administered Quality of Life Inventory (QOLI) in different psychiatric conditions. *Cognitive Behaviour Therapy*, *42*(4), 1–13.

Lindsay, D. (1988). Assessing precision in the manuscript review process: A little better than a dice role. *Sociometrics*, *14*, 75–82.

Littell, R. C. (2013). *SAS: Statistical and numerical computing*. Published online, John Wiley & Sons.

Litten, R. Z., Ryan, M. L., Fertig, J. B., Falk, D. E., Johnson, B., Dunn, K. E., . . . Stout, R. (2013). A double-blind, placebo-controlled trial assessing the efficacy of varenicline tartrate for alcohol dependence. *Journal of Addiction*, *7*(4), 277–286.

Little, T. L. (Ed.) (2013). *The Oxford handbook of quantitative methods* (Vols. 1 & 2). New York: Oxford University Press.

Littner, Y., Mimouni, F. B., Dollberg, S., & Mandel, D. (2005). Negative results and impact factor: A lesson from neonatology. *Archives of Pediatric and Adolescent Medicine*, *159*, 1036–1037.

Lucas, R. E., & Diener, E. (2009). Personality and subjective well-being. In E. Diener (Ed.), *The science of well-being* (pp. 75–102). Netherlands: Springer.

Lund, H., Snilsberg, A. H., Paus, E., Halvorsen, T. G., Hemmersbach, P., & Reubsaet, L. (2013). Sports drug testing using immuno-MS: Clinical study comprising administration of human chorionic gonadotropin to males. *Analytical and bioanalytical chemistry*, *405*(5), 1569–1576.

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Anna Rottger, M., Jorasz, C., . . . Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, *23*, 14–24.

Lutz, W., Stulz, N., & Kock, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of Affective Disorders*, *118*, 60–68.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

Maarse, F. J., Mulder, L. J. M., Brand, A. N., & Akkerman, A. E. (Eds.) (2003). *Clinical assessment, computerized methods, and instrumentation*. Lisse, The Netherlands: Swetz & Zeitlinger, B.V.

MacDonald, J. M., Morral, A., & Piquero, A. R. (2011). Socially desirable response bias in criminology: An example of its effect in testing the effects of self-control. *Measuring Crime & Criminality. Advances in Criminological Theory*, *17*, 21–36.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS [management information systems] and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*, 293–334.

Madden, G. J. (Ed.). (2013). *APA handbook of behavior analysis* (Volumes 1 & 2). Washington, DC: American Psychological Association.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Malakoff, D. (2013). Hey, you've got to hide your work away. *Science*, *342*, 70–71.

Malakoff, D., & Enserink, M. (2013). Dual use research: New U.S. rules increase oversight of H5N1 studies, other risky science. *Science*, *339*, 1025.

Maliken, A. C., & Katz, L. F. (2013). Exploring the impact of parental psychopathology and emotion regulation on evidence-based parenting interventions: A transdiagnostic approach to improving treatment effectiveness. *Clinical Child and Family Psychology Review*, 1–14. Available at http://link.springer.com/article/10.1007/s10567-013-0132-4#page-1

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32,* 860–875.

Manolov, R., & Solanas, A. (2009). Problems of the randomization tests for AB designs. *Psicologica, 30,* 137–154.

Mansergh, G., Koblin, B. A., McKirnan, D. J., Hudson, S. M., Flores, S. A., Wiegand, R. E., . . . Colfax, G. N. (2010). An intervention to reduce HIV risk behavior of substance-using men who have sex with men: A two-group randomized trial with a nonrandomized third group. *PLoS Medicine*, *7*(8), e1000329.

Manson, S. M., Garroutte, E., Goins, R. T., & Henderson, P. N. (2004). Access, relevance, and control in the research process: Lessons from Indian country. *Journal of Aging and Health*, *16*(5 suppl), 58S–77S.

Manson, S. M. (Ed.) (1989). *American Indian and Alaska Native Mental Health Research*, *2* (whole issue number 3).

Marciano, P. L., & Kazdin, A. E. (1994). Self-esteem, depression, hopelessness, and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Clinical Child Psychology*, *23*, 151–160.

Marker, C. D., Comer, J. S., Abramova, V., & Kendall, P. C. (2013). The reciprocal relationship between alliance and symptom improvement across the treatment of childhood anxiety. *Journal of Clinical Child and Adolescent Psychology*, *42*, 22–33.

Marušić, A., Bošnjak, L., & Jerončić, A. (2011). A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. *Plos one*, *6*(9), e23477.

Masdeu, J. C. (2011). Neuroimaging in psychiatric disorders. *Neurotherapeutics*, *8*(1), 93–102.

Matthews, N., Spears, L., & Ball, J. (2012). Bathroom BANTER: Sex, love, and the bathroom wall. *Electronic Journal of Human Sexuality*, *15*, www.ejhs.org

Maulik, P. K., Eaton, W. W., & Bradshaw, C. P. (2010). The effect of social networks and social support on common mental disorders following specific life events. *Acta Psychiatrica Scandinavica*, *122*(2), 118–128.

Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry*, *18*(8), 655–661.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research. *Psychological Methods*, *9*, 147–163.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York City: Houghton Mifflin Harcourt.

McCabe, R., & Priebe, S. (2004). The therapeutic relationship in the treatment of severe mental illness: A review of methods and findings. *International Journal of Social Psychiatry*, *50*, 115–128.

McCambridge, J., Butor-Bhavsar, K., Witton, J., & Elbourne, D. (2011). Can research assessments themselves cause bias in behaviour change trials? A systematic review of evidence from Solomon 4-Group studies. *PLoS One*, *6*(10), e25223.

McCleary, R. M., & McDowall, D. (2012). Time-series designs. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 2, pp. 613–627). Washington, DC: American Psychological Association.

McCrae, R. R., & Costa, P. T. (2010). *NEO Inventories for the NEO Personality Inventory-3 (NEO PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3) and NEO Personality Inventory-revised (NEO PI-R): Professional Manual*. Lutz, FL: Psychological Assessment Resources.

McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., . . . Mayberg, H. S. (2013). Toward a neuroimaging treatment selection biomarker for major depressive disordertreatment-specific biomarker for major depressiontreatment-specific biomarker for major depression. *JAMA psychiatry*, *70*(8), 821–829.

McGrath, R. E., & Carroll, E. J. (2012). The current status of "projective" "tests." In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 329–348). Washington, DC: American Psychological Association.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, *48*, 1–30.

McLeod, J. (2011). *Qualitative research in counselling and psychotherapy* (2nd ed.). London: Sage.

McNulty, J. K., Olson, M. A., Meltzer, A. L., & Shaffer, M. J. (2013). Though they may be unaware, newlyweds implicitly know whether their marriage will be satisfying. *Science*, *342*(6162), 1119–1120.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Meijer, A., Zuidersma, M., & de Jonge, P. (2013). Depression as a non-causal variable risk marker in coronary heart disease. *BioMedCentral Medicine*, *11*(1), 130. Available at www.biomedcentral.com/1741-7015/11/130

Mennes, M., Biswal, B., Castellanos, F. X., & Milham, M. P. (2013). Making data sharing work: The FCP/INDI experience. *Neuroimage*, *82*, 683–691.

Merry, S. N., Stasiak, K., Shepherd, M., Frampton, C., Fleming, T., & Lucassen, M. F. (2012). The effectiveness of SPARX, a computerised self-help intervention for adolescents seeking help for depression: Randomised controlled non-inferiority trial. *British Medical Journal*, *344*, e2598. Available on line at www.ncbi.nlm.nih.gov/pmc/articles/PMC3330131/

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, & Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist*, *56*, 128–165.

Meyer-Lindenberg, A., Buckholtz, J. W., Kolachana, B., Hariri, A. R., Pezawas, L., Blasi, G., . . . Weinberger, D. R. (2006). Neural mechanisms of genetic risk for impulsivity and violence in humans. *Proceedings of the National Academy of Sciences*, *103*, 6269–6274.

Mezzich, J. E. (1995). Cultural formulation and comprehensive diagnosis: Clinical and research perspectives. *Psychiatric Clinics of North America*, *18*, 649–657.

Michelson, D., Davenport, C., Dretzke, J., Barlow, J., & Day, C. (2013). Do evidence-based interventions work when tested in the "real world?" A systematic review and meta-analysis of parent management training for the treatment of child disruptive behavior. *Clinical Child and Family Psychology Review*, *16*(1), 18–34.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: Sage.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378.

Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.

Miller, D. J., & Hersen, M. (Eds.) (1992). *Research fraud in the behavioral and biomedical sciences*. New York: John Wiley & Sons.

Miller, G. E., & Chen, E. (2010). Harsh family climate in early life presages the emergence of a proinflammatory phenotype in adolescence. *Psychological Science, 21*(6), 848–856.

Miller, L., & Reynolds, J. (2009). Autism and vaccination—the current evidence. *Journal for Specialists in Pediatric Nursing*, *14*(3), 166–172.

Miller, L. R., & Das, S. K. (2007). Cigarette smoking and Parkinson's disease. *Experimental and Clinical Sciences International*, *6*, 93–99.

Miller, W. E., & Kreiner, D. S. (2008). Student perception of coercion to participate in psychological research. *North American Journal of Psychology*, *10*, 53–64.

Miltenberger, R. G., Flessner, C., Gatheridge, B., Johnson, B., Satterlund, M., & Egemo, K. (2004). Evaluation of behavioral skills training to prevent gun play in children. *Journal of Applied Behavior Analysis, 37*, 513–516.

Mimouni-Bloch, A., Kachevanskaya, A., Mimouni, F. B., Shuper, A., Raveh, E., & Linder, N. (2013). Breastfeeding may protect from developing attention-deficit/hyperactivity disorder. *Breastfeeding Medicine*. Available at http://online.liebertpub.com/doi/abs/10.1089/bfm.2012.0145

Mischel, W., & Ayduk, O. (2002). Self-regulation in a cognitive—affective personality system: Attentional control in the service of the self. *Self and Identity*, *1*(2), 113–120.

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., . . . Shoda, Y. (2011). "Willpower" over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256.

Mitchell, M. D., Gehrman, P., Perlis, M., & Umscheid, C. A. (2012). Comparative effectiveness of cognitive behavioral therapy for insomnia: A systematic review. *BMC Family Practice*, *13*(1), 40.

Mitchell, M. L., & Jolley, J. M. (2012). *Research design explained* (7th ed.). Belmont, CA: Wadsworth.

Moffitt, T. E. (2005). The new look of behavioral genetics in developmental psychopathology: Gene-environment interplay in antisocial behaviors. *Psychological Bulletin*, *131*, 533–554.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693–2698.

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., & Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine*, *40*, 899–909.

Moher, D., Schulz, K. F., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of American Medical Association*, *285*, 1987–1991.

Mohr, L. B. (1990). *Understanding significance testing.* Newbury Park, CA: Sage.

Molenaar, P. J., Boom, Y., Peen, J., Schoevers, R. A., Van, R., & Dekker, J. J. (2011). Is there a dose–effect relationship between the number of psychotherapy sessions and improvement of social functioning? *British Journal of Clinical Psychology*, *50*(3), 268–282.

Moncrieff, J., Wessely, S., & Hardy, R. (2004). Active placebos versus antidepressants for depression. *Cochrane Database Systematic Reviews*, 1, CD003012.

Montag, C., Ehrlich, A., Neuhaus, K., Dziobek, I., Heekeren, H. R., Heinz, A., & Gallinat, J. (2010). Theory of mind impairments in euthymic bipolar patients. *Journal of Affective Disorders*, *123*(1), 264–269.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Medicine*, *4*(2): e28.

Moore, R. A., Derry, S., & McQuay, H. J. (2010). Fraud or flawed: Adverse impact of fabricated or poor quality research. *Anaesthesia*, *65*(4), 327–330.

Morell, V. (2013, February 20). Dolphins can call each other, not by name, but by whistle. *Science Now*. Available at http://news.sciencemag.org/plants-animals/2013/02/dolphins-can-call-each-other-not-name-whistle

Morell, V. (2014, February 18). Elephants console each other. *Science Now.* Available at http://news.sciencemag.org/plants-animals/2014/02/elephants-console-each-other

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2012). Response strategies and response styles in cross-cultural surveys. *Cross-Cultural Research*, *46*(3), 255–279.

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2013). Exploring the response process of culturally differing survey respondents with a response style: A sequential mixed methods study. *Field Methods*, *25*(2), 162–181.

Morris, B. (2005). *Discovering bits and pieces of me: Research exploring women's experiences of psychoanalytic psychotherapy*. London: Women's Therapy Centre.

Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., & Deleeuw, W. (2010). Mobile therapy: Case study evaluations of a cell phone application for emotional self-awareness. *Journal of Medical Internet Research*,*12*(2), e10.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., . . . Wray, N. P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New England Journal of Medicine*, *347*(2), 81–88.

Mosteller, F. (2010). *The pleasure of statistics: The autobiography of Frederick Mosteller*. S.E. Feinberg, D.C. Hoaglin, & J.M. Tanur (Eds.). New York: Springer.

Moyer, A., & Franklin, N. (2011). Strengthening the educational value of undergraduate participation in research as part of a psychology department subject pool. *Journal of Empirical Research on Human Research Ethics: An International Journal*, *6*(1), 75–82.

Mueller, M. M., Moore, J. W., Doggett, R. A., & Tingstrom, D. H. (2000). The effectiveness of contingency-specific and contingency-nonspecific prompts in controlling bathroom graffiti. *Journal of Applied Behavior Analysis*, *33*(1), 89–92.

Mulcahy, R., Reay, R. E., Wilkinson, R. B., & Owen, C. (2010). A randomised control trial for the effectiveness of group interpersonal psychotherapy for postnatal depression. *Archives of Women's Mental Health*, *13*(2), 125–139.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*, 852–863.

Müller, J. M., Postert, C., Beyer, T., Furniss, T., & Achtergarde, S. (2010). Comparison of eleven short versions of the Symptom Checklist 90-Revised (SCL-90-R) for use in the assessment of general psychopathology. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 246–254.

Multiple Risk Factor Intervention Trial Research Group. (1982). Multiple risk factor intervention trial. *Journal of the American Medical Association*, *248*(12), 1465–1477.

Murphy, K. R., Myors, B., & Wolach, A. H. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (3rd ed.). New York: Routledge.

Murphy, M. L., & Pichichero, M. E. (2002). Prospective identification and treatment of children with pediatric autoimmune neuropsychiatric disorder associated with group A streptococcal infection (PANDAS). *Archives of Pediatrics & Adolescent Medicine*, *156*(4), 356–361.

Murray-Close, D., Ostrov, J. M., Nelson, D. A., Crick, N. R., & Coccaro, E. F. (2010). Proactive, reactive, and romantic relational aggression in adulthood: Measurement, predictive validity, gender differences, and association with intermittent explosive disorder. *Journal of Psychiatric Research*, *44*(6), 393–404.

Musser, E. H., Bray, M. A., Kehle, T. J., & Jenson, W. R. (2001). Reducing disruptive behaviors in students with serious emotional disturbance. *School Psychology Review*, *30*, 294–304.

Nathan, P. E., & Gorman, J. M. (Eds.) (2015). *Treatments that work* (4th ed.). New York: Oxford University Press.

National Aeronautics and Space Administration (2009, April). NASA's Great Observatories. Available at www.nasa.gov/audience/forstudents/postsecondary/features/F_NASA_Great_Observatories_PS.html

National Institutes of Health. (2010). National Institute of Neurological Disorders and Stroke. Grand opportunities in comparative effectiveness research. Available at www.ninds.nih.gov/recovery/arra-funding/go-cer.htm

National Institutes of Health. (2011). Bioethics resources on the web. Available at www.nih.gov/sigs/bioethics/conflict.html.

National Institutes of Health. (2013a). Certificate of confidentiality. Available at http://grants.nih.gov/grants/policy/coc/

National Institutes of Health. (2013b). Comparative effectiveness research. National Information Center on Health Services Research and Health Care Technology. Available at www.nlm.nih.gov/hsrinfo/cer.html

National Institutes of Health. (2013c). Daily-use HIV prevention approaches prove ineffective among women in NIH study. Available at www.nih.gov/news/health/mar2013/niaid-04a.htm

National Institutes of Health. (2013d). NIH commits $24 million annually for big data centers of excellence. Available at www.nih.gov/news/health/jul2013/nih-22.htm

National Institutes of Health. (2013e). NIH policy on mitigating risks of life sciences dual use research of concern. Notice Number: OT-OD-13-107. Bethesda, MD: NIH. Available at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-107.html

National Institutes of Health. (2013f). NIH to reduce significantly the use of chimpanzees in research. Available at www.nih.gov/news/health/jun2013/od-26.htm

National Institute of Mental Health. (2013g). Questions and answers about the National Comorbidity Survey Replication (NCSR) Study. Available at www.nimh.nih.gov/health/topics/statistics/ncsr-study/questions-and-answers-about-the-national-comorbidity-survey-replication-ncsr-study.shtml

National Research Council. (2001). *Frontiers of engineering: Reports on leading-edge engineering from the 2000 NAE Symposium on Frontiers in Engineering*. Washington, DC: The National Academies Press.

National Research Council. (2002). Integrity in scientific research: *Creating an environment that promotes responsible conduct*. Washington, DC: National Academy Press. http://books.nap.edu/catalog.php?record_id=10430#toc

National Research Council. (2004). Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology. *Biotechnology research in an age of terrorism*. Washington, DC: National Academy Press.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. Panel on Handling Missing Data in Clinical Trials. Washington, DC: National Academies Press. Available at www.nap.edu/catalog.php?record_id=12955

National Research Council. (2013). *Frontiers in massive data analysis.* Washington, DC: The National Academies Press.

Navajo Nation. (2009). Navajo Nation Human Research Review Board: PI Guidelines. Retrieved October 2013 from www.nnhrrb.navajo-nsn.gov/pdf/First%20Procedural%20Guidelines%20for%20PI%20_2_.pdf

Neale, B., Henwood, K., & Holland, J. (2012). Researching lives through time: An introduction to the Timescapes approach. *Qualitative Research*, *12*(1), 4–15.

Needleman, H. L., & Bellinger, D. (1984). The developmental consequences of childhood exposure to lead: Recent studies and methodological issues. In B.B. Lahey & A.E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 7, pp. 195–220). New York: Plenum.

Needleman, H. L., Schell, A. S., Bellinger, D., Leviton, A., & Alldred, E. N. (1990). The long-term effects of exposure to low doses of lead in childhood: An 11-year follow-up report. *New England Journal of Medicine*, *322*, 83.

Nelson, C., & Sheridan, M. (2011). Lessons from neuroscience research for understanding causal links between family and neighborhood characteristics and educational outcomes. In G. Duncan & R. Murname (Eds.), *Whither opportunity: Rethinking the role of neighborhoods and families on schools and school outcomes for American children* (pp. 27–46). New York: Russell Sage Foundation.

Nelson, J. C., & Devanand, D. P. (2011). A systematic review and meta-analysis of placebo-controlled antidepressant studies in people with depression and dementia. *Journal of the American Geriatrics Society*, *59*(4), 577–585.

Nettleton, S., Neale, J., & Stevenson, C. (2012). Sleeping at the margins: A qualitative study of homeless drug users who stay in emergency hostels and shelters. *Critical Public Health*, *22*(3), 319–332.

Neufeld, E., O'Rourke, N., & Donnelly, M. (2010). Enhanced measurement sensitivity of hopeless ideation among older adults at risk of self-harm: Reliability and validity of Likert-type responses to the Beck Hopelessness Scale. *Aging & Mental Health*, *14*(6), 752–756.

New York Academy of Sciences (2010). The biology of disadvantage: Socioeconomic status and health, *Annals of the New York Academy of Sciences*, *1186*, 1–275. Special issue available online at http://onlinelibrary.wiley.com/doi/10.1111/nyas.2010.1186.issue-1/issuetoc

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *294*, 175–240 (Part 1), 263–294 (Part 2).

Ngai, S., Gold, J. L., Gill, S. S., & Rochon, P. A. (2005). Haunted manuscripts: Ghost authorship in the medical literature. *Accountability in Research: Policies and Quality Assurance*, *12*(2), 103–114.

Nichols, D. S. (2011). *Essentials of MMPI-2 assessment* (Vol. 88). New York: John Wiley & Sons.

Nielsen, L., & Burlingame, N. (2012). *A simple introduction to data science*. Wickford, RI: New Street Communications.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Nilsson, T., Svensson, M., Sandell, R., & Clinton, D. (2007). Patients' experiences of change in cognitive–behavioral therapy and psychodynamic therapy: A qualitative comparative study. *Psychotherapy Research*, *17*(5), 553–566.

Nock, M. K., & Banaji, M. R. (2007). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology*, *75*(5), 707–715.

Nock, M. K., Deming, C. A., Fullerton, C. S., Gilman, S. E., Goldenberg, M., Kessler, R. C., . . . Ursano, R. J. (2013). Suicide among soldiers: A review of psychosocial risk and protective factors. *Psychiatry: Interpersonal & Biological Processes*, *76*(2), 97–125.

Nock, M. K., & Kazdin, A. E. (2005). Randomized controlled trial of a brief intervention for increasing participation in parent management training. *Journal of Consulting and Clinical Psychology*, *75*, 872–879.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirksy, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*, 400–424.

Norcross, J. C. (Ed.). (2011). *Psychotherapy relationships that work: Evidence-based responsiveness* (2nd ed.). New York: Oxford University Press.

Nordin, S., Carlbring, P., Cuijpers, P., & Andersson, G. (2010). Expanding the limits of bibliotherapy for panic disorder: Randomized trial of self-help without support but with a clear deadline. *Behavior Therapy*, *41*(3), 267–276.

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification*, *39*, 295–314.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Noussair, C., Robin, S., & Ruffieux, B. (2004). Do consumers really refuse to buy genetically modified food? *The Economic Journal*, *114*(492), 102–120.

O'Connell-Rodwell, C. E., Arnason, B., & Hart, L. A. (2000). Seismic properties of elephant vocalizations and locomotion. *Journal of the Acoustical Society of America*, *108*, 3066.

Ogles, B. M. (2013). Measuring change in psychotherapy research. In M.J. Lambert (Ed.), *Bergin & Garfield's Handbook of psychotherapy and behavior change* (6th ed., pp. 134–167). Hoboken, NJ: John Wiley & Sons.

Okuyama, T., Yokoi, S., Abe, H., Isoe, Y., Suehiro, Y., Imada, H., . . . Takeuchi, H. (2014). A neural mechanism underlying mating

preferences for familiar individuals in medaka fish. *Science*, *343*(6166), 91–94.

Olatunji, B. O., Davis, M. L., Powers, M. B., & Smits, J. A. (2013). Cognitive-behavioral therapy for obsessive-compulsive disorder: A meta-analysis of treatment outcome and moderators. *Journal of Psychiatric Research*, *47*, 33–41.

O'Neil, A., Sanderson, K., Oldenburg, B., & Taylor, C. B. (2011). Impact of depression treatment on mental and physical health-related quality of life of cardiac patients: A meta-analysis. *Journal of Cardiopulmonary Rehabilitation and Prevention*, *31*(3), 146–156.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality & Quantity*, *41*(2), 233–249.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Opie, L. H., & Lecour, S. (2007). The red wine hypothesis: from concepts to protective signalling molecules. *European Heart Journal*, *28*(14), 1683–1693.

Orenstein, W. A. (2013). Efforts on track to eradicate world of polio by 2018. *American Academy of Pediatrics News*. Available at http://aapnews.aappublications.org/content/early/2013/06/03/aapnews.20130603-1

Orlinsky, D. E., RØnnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M.J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 307–389). New York: John Wiley & Sons.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Orne, M. T., & Scheibe, K. E. (1964). The contribution of nondeprivation factors in the production of sensory deprivation effects: The psychology of the "panic button." *Journal of Abnormal and Social Psychology*, *68*, 3–12.

Osher, Y., Dobron, A., Belmaker, R. H., Bersudsky, Y., & Dwolatzky, T. (2011). Computerized testing of neurocognitive function in euthymic bipolar patients compared to those with mild cognitive impairment and cognitively healthy controls. *Psychotherapy and Psychosomatics*, *80*(5), 298–303.

Osrin, D., Azad, K., Fernandez, A., Manandhar, D. S., Mwansambo, C.W., Tripathy, P., & Costello, A. M. (2009). Ethical challenges in cluster randomized controlled trials: Experiences from public health interventions in Africa and Asia. *Bulletin of the World Health Organization*, *87*, 772–779. Available at www.scielosp.org/scielo.php?pid=S0042-96862009001000013&script=sci_arttext

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *l05*(2), 171–192.

Owens, C., & White, F. A. (2013). A 5-year systematic strategy to reduce plagiarism among first-year psychology university students. *Australian Journal of Psychology*, *65*(1), 14–21.

Paajanen, T. A., Oksala, N. K., Kuukasjärvi, P., & Karhunen, P. J. (2010). Short stature is associated with coronary heart disease: A systematic review of the literature and a meta-analysis. *European Heart Journal*, *31*(14), 1802–1809.

Packer, M. J. (2011). *The science of qualitative research*. New York: Cambridge University Press.

Palmour, N., Affleck, W., Bell, E., Deslauriers, C., Pike, B., Doyon, J., & Racine, E. (2011). Informed consent for MRI and fMRI research: Analysis of a sample of Canadian consent documents. *BMC Medical Ethics*, *12*(1), 1.

Palta, M., Prineas, R. J., Berman, R., & Hannan, P. (1982). Comparison of self-reported and measured height and weight. *American Journal of Epidemiology*, *115*(2), 223–230.

Paniagua, F. A., & Yamada, A. (Eds.). (2013). *Handbook of multicultural mental health: Assessment and treatment of diverse populations* (2nd ed.). San Diego, CA: Academic Press.

Pappworth, M. H. (1967). *Human guinea pigs: Experimentation on man*. London: Routledge & K. Paul.

Park., H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis of single-case designs. *Journal of Experimental Education*, *58*, 311–320.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods, *Behavior Therapy, 34,* 189–211.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly*, *21*, 418–443.

Parker, R. I., & Hagan-Burke, S. (2007a). Single case research results as clinical outcomes. *Journal of School Psychology*, *45*, 637–653.

Parker, R. I., & Hagan-Burke, S. (2007b). Useful effect size interpretations for single case research. *Behavior Therapy*, *38*, 95–105.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative items for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). New York: Springer.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., . . . Kirkwood, B. R. (2010). Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): A cluster randomised controlled trial. *Lancet*, *376*, 2086–2095.

Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., . . . Kirkwood, B. R. (2011). Lay health worker led intervention for depressive and anxiety disorders in India: Impact on clinical and disability outcomes over 12 months. *British Journal of Psychiatry, 199*, 459–466.

Pattar, U., Raybagkar, V. H., & Garg, S. (2012). Teaching-Learning through innovative experiments: An investigation of students' responses. *Latin-American Journal of Physics Education, 6,* 347–352.

Patterson, T. L., & Mausbach, B. T. (2010). Measurement of functional capacity: A new approach to understanding functional differences and real-world behavioral adaptation in those with mental illness. *Annual Review of Clinical Psychology*, *6*, 139–154.

Paul, G. L. (1966). *Insight versus desensitization in psychotherapy: An experiment in anxiety reduction*. Stanford, CA: Stanford University Press.

Paul, G. L. (1967). Outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*, 109–118.

Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology*, *103*(1), 158–175.

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, *85*(1), 193–202.

Pazda, A. D., Elliot, A. J., & Greitemeyer, T. (2012). Sexy red: Perceived sexual receptivity mediates the red-attraction relation in men viewing woman. *Journal of Experimental Social Psychology*, *48*(3), 787–790.

Peng, C. Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education*, *80*(2), 113–136.

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77,* 212–218.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, *75,* 829–841.

Perkel, J. M. (2013, January 18). Life science technologies: This is your brain: mapping the connectome science. *Science*, DOI: 10.1126/science.opms.p1300071. Retrieved February 2, 2013, from www.sciencemag.org/site/products/lst_20130118.xhtml

Perry, G. (2012). *Behind the shock machine: The untold story of the notorious Milgram psychology experiments*. New York: The New Press.

Petersen, S., Hydeman, J., & Flowers, K. (2011). The Decisional Processing Model: How cognitive processing affects adherence to mammography among African American Women. *Journal of Black Psychology*, *37*(3), 357–380.

Peterson, L., Tremblay, G., Ewigman, B., & Popkey, C. (2002). The Parental Daily Diary: A sensitive measure of the process of change in a child maltreatment prevention program. *Behavior Modification*, *26*, 594–604.

Peterson, M. D., Rhea, M. R., Sen, A., & Gordon, P. M. (2010). Resistance exercise for muscular strength in older adults: A meta-analysis. *Ageing research reviews*, *9*(3), 226–237.

Pfeffer, C., & Olsen, B. R. (2002). Editorial: Journal of negative results in biomedicine. *Journal of Negative Results in Biomedicine*, *1*(1), 2.

Pharoah, F., Mari, J., Rathbone, J., & Wong, W. (2010). Family intervention for schizophrenia. *Cochrane Database of Systematic Reviews*, *12*. Available at http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000088.pub3/abstract

Phillips, T. M., Randall, B. A., Peterson, D. J., Wilmoth, J. D., & Pickering, L. E. (2013). Personal problems among rural youth and their relation to psychosocial well-being. *Journal of Extension*, *51*(3), 3FEA9.

Pistrang, N., & Barker, C. (2012). Varieties of qualitative research: A pragmatic approach to selecting methods. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 2, pp. 5–18). Washington, DC: American Psychological Association.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS one*, *2*(3), e308.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569.

Poduska, J. M., & Bowes, S. (April 2010). Good Behavior Game: A classroom behavior management strategy. Presented at Blueprints for Violence Prevention, San Antonio, Texas. Available at www.blueprintsconference.com/2010/presentations/f7a_jp.pdf

Pohl, R. F. (Ed.) (2012). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory*. New York: Psychology Press.

Porto, P. R., Oliveira, L., Mari, J., Volchan, E., Figueira, I., & Ventura, P. (2009). Does cognitive behavioral therapy change the brain? A systematic review of neuroimaging in anxiety disorders. *Journal of Neuropsychiatry and Clinical Neurosciences*, *21*, 114–125.

Portzky, G., & van Heeringen, K. (2006). Suicide prevention in adolescents: A controlled study of the effectiveness of a school-based psycho-educational program. *Journal of Child Psychology and Psychiatry, 47*, 910–918.

Power, C., Kuh, D., & Morton, S. (2013). From developmental origins of adult disease to life course research on adult disease and aging: Insights from birth cohort studies. *Public Health*, *34*(1), 7–28.

Powers, M. B., Halpern, J. M., Ferenschak, M. P., Gillihan, S. J., & Foa, E. B. (2010). A meta-analytic review of prolonged exposure for posttraumatic stress disorder. *Clinical Psychology Review*, *30*(6), 635–641.

Poythress, N., Epstein, M., Stiles, P., & Edens, J. F. (2011). Awareness of the tuskegee syphilis study: Impact on offenders' decisions to decline research participation. *Behavioral Sciences & the Law*, *29*(6), 821–828.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

Prinstein, M. J. (Ed.) (2013). *The portable mentor: Expert guide to a successful career in psychology* (2nd ed.). New York: Springer Science-Business Media.

Prinz, U., Nutzinger, D. O., Schulz, H., Petermann, F., Braukhaus, C., & Andreas, S. (2013). Comparative psychometric analyses of the SCL-90-R and its short versions in patients with affective disorders. *BMC Psychiatry*, *13*(1), 104. Available at www.biomedcentral.com/content/pdf/1471-244X-13-104.pdf

Prochaska, J. O., & Norcross, J. C. (2010). Prochaska, J. O., & Norcross, J. C. (2010). *Systems of psychotherapy: A transtheoretical analysis* (7th ed.). Belmont, CA: Brooks Cole, Cengage Learning.

Protzko, J., Aronson, J., & Blair, C. (2013). How to make a young child smarter evidence from the database of raising intelligence. *Perspectives on Psychological Science*, *8*(1), 25–40.

Quesnel, C., Savard, J., Simard, S., Ivers, H., & Morin, C. M. (2003). Efficacy of cognitive-behavioral therapy for insomnia in women treated for nonmetastatic breast cancer. *Journal of Consulting and Clinical Psychology*, *71*, 189–200.

Quidé, Y., Witteveen, A. B., El-Hage, W., Veltman, D. J., & Olff, M. (2012). Differences between effects of psychological versus pharmacological treatments on functional and morphological brain alterations in anxiety disorders and major depressive disorder: A systematic review. *Neuroscience and Biobehavioral Reviews*, *36*, 626–644.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice*, *31*(4), 2–12.

Rathgeb-Fuetsch, M., Kempter, G., Feil, A., Pollmächer, T., & Schuld, A. (2011). Short- and long-term efficacy of cognitive behavioral therapy for DSM-IV panic disorder in patients with and without severe psychiatric comorbidity. *Journal of Psychiatric Research*, *45*(9), 1264–1268.

Rathi, V., Dzara, K., Gross, C. P., Hrynaszkiewicz, I., Joffe, S., Krumholz, H. M., . . . Ross, J. S. (2012). Sharing of clinical trial data among trialists: A cross-sectional survey. *British Medical Journal*, *345,* e7570.

Rau, T. J., Merrill, L. L., McWhorter, S. K., Stander, V. A., Thomsen, C. J., Dyslin, C.W., . . . Milner, J. (2011). Evaluation of a sexual assault education/prevention program for female U.S. Navy personnel. *Military Medicine*, *176*(10), 1178–1183.

Rebollo, R., Horard, B., Hubert, B., & Vieira, C. (2010). Jumping genes and epigenetics: Towards new species. *Gene*, *454*(1), 1–7.

Resnick, J. H., & Schwartz, T. (1973). Ethical standards as an independent variable in psychological research. *American Psychologist*, *28*, 134–139.

Resnick, P. A., Galovski, T. E., Uhlmansiek, M. O. B., Scher, C. D., Clum, G. A., & Young-Xu, Y. (2008). A randomized clinical trial to dismantle components of cognitive processing therapy for posttraumatic stress disorder in female victims of interpersonal violence. *Journal of Consulting and Clinical Psychology*, *76*(2), 243–258.

Reyes, J. R., Vollmer, T. R., Sloman, K. N., Hall, A., Reed, R., Jansen, G., . . . Stoutimore, M. (2006). Assessment of deviant arousal in adult male sex offenders with developmental disabilities. *Journal of Applied Behavior Analysis, 39*, 173–188.

Rice, V. H., & Stead, L. F. (2008). Nursing interventions for smoking cessation. *Cochrane Database of Systematic Reviews,* Issue 1 (Art. No. CD001188).

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147.

Roebuck, K. (2012). *Caqdas - Computer assisted/aided qualitative data analysis*. Newstead, Queensland, Australia: Emereo Publishing.

Roediger, H. L., III, & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science*, *9*, 123–127.

Roffman, J. L., Marci, C. M., Glick, D. M., Dougherty, D. D., & Rauch, S. L. (2005). Neuroimaging and the functional neuroanatomy of psychotherapy. *Psychological Medicine, 35*, 1385–1398.

Rogers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1–12.

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality, 47*(4), 330–338.

Rolls, G. (2010). *Classic case studies in psychology*. London: Hodder Education.

Ronk, F. R., Hooke, G. R., & Page, A. C. (2012). How consistent are clinical significance classifications when calculation methods and outcome measures differ? *Clinical Psychology: Science and Practice*, *19*(2), 167–179.

Rosenbaum, J. E. (2009). Patient teenagers? A comparison of the sexual behavior of virginity pledgers and matched nonpledgers. *Pediatrics, 123*, e110–e120.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer-Verlag.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged edition). New York: Irvington.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1991). Replication in behavioral research. In J.W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–30). Newbury Park, CA: Sage.

Rosenthal, R., & Rosnow, R. (2007). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York: McGraw Hill.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, *105*, 143–146.

Ross, J. S., Hill, K. P., Egilman, D. S., & Krumholz, H. M. (2008). Guest authorship and ghostwriting in publications related to Rofecoxib: A case study of industry documents from Rofecoxib litigation. *Journal of the American Medical Association*, *299*, 1800–1812.

Ross, L. F., Loup, A., Nelson, R. M., Botkin, J. R., Kost, R., Smith Jr, G. R., & Gehlert, S. (2010). Human subjects protections in community-engaged research: A research ethics framework. *Journal of Empirical Research on Human Research Ethics*, *5*(1), 5–17.

Roth, A., & Fonagy, P. (2005). *What works for whom?: A critical review of psychotherapy research* (2nd ed.). New York: Guilford Press.

Roth, L. W., & Polotsky, A. J. (2012). Can we live longer by eating less? A review of caloric restriction and longevity. *Maturitas*, *71*, 315–319.

Rothenstein, J. M., Tomlinson, G., Tannock, I. F., & Detsky, A. F. (2011). Company stock prices before and after public announcements related to oncology drugs. *Journal of the National Cancer Institute, 103,* 1507–1512.

Roukos, D. H. (2009). Twenty-one–gene assay: Challenges and promises in translating personal genomics and whole-genome scans into personalized treatment of breast cancer. *Journal of Clinical Oncology*, *27*(8), 1337–1338.

Rubel, S. K., Miller, J. W., Stephens, R. L., Xu, Y., Scholl, L. E., Holden, E. W., . . . Volk, R. J. (2010). Testing the effects of a decision aid for prostate cancer screening. *Journal of Health Communication*, *15*(3), 307–321.

Rumbaugh, K. P. (Ed.) (2011). *Quorum sensing: Methods and protocols*. New York City: Humana Press.

Russell, N. J. C. (2011). Milgram's obedience to authority experiments: Origins and early evolution. *British Journal of Social Psychology*, *50,* 140–162. doi:10.1348/014466610X492205

Rutledge, T., & Loh, C. (2004). Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine*, *27*(2), 138–145.

Rutter, M. B. (1981). Epidemiological/longitudinal strategies and causal research in child psychiatry. *Journal of the American Academy of Child Psychiatry*, *20*, 513–544.

Rutter, M. B., Chadwick, O., & Shaffer, D. (1983). Head injury. In M.B. Rutter (Ed.), *Developmental neuropsychiatry* (pp. 83–111). New York: Guilford.

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, *1*, 114–117.

Rylands, A. J., McKie, S., Elliott, R., Deakin, J. W., & Tarrier, N. (2011). A functional magnetic resonance imaging paradigm of expressed emotion in schizophrenia. *Journal of Nervous and Mental Disease*, *199*(1), 25–29.

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, *24*(5), 563–572.

Samaan, Z., Anand, S., Zhang, X., Desai, D., Rivera, M., Pare, G., . . . Meyre, D. (2013). The protective effect of the obesity-associated rs9939609: A variant in fat mass- and obesity-associated gene on depression. *Molecular Psychiatry, 18,* 1281–1286.

Sandelowski, M. (2012). Metasynthesis of qualitative research. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 2, pp. 19–36). Washington, DC: American Psychological Association.

Sandøe, P. (2013). *Ethics of animal use*. New York: John Wiley & Sons.

Santangelo, P., Ebner-Priemer, U. W., & Trull, T. J. (2013). Experience sampling methods of clinical psychology. In J.S. Comer & P.C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 188–209). New York: Oxford University Press.

Satcher, D. (2001). Department of Health and Human Services (2001). *Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the Surgeon General.* Washington, DC: U.S. Department of Health and Human Services.

Sawilowsky, S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. *The Journal of Experimental Education*, *62*(4), 361–376.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*(1), 91–106.

Schaufeli, W. B., Shimazu, A., & Taris, T. W. (2009). Being driven to work excessively hard: The evaluation of a two-factor measure of workaholism in the Netherlands and Japan. *Cross-Cultural Research*, *43*(4), 320–348.

Schenker, Y., Fernandez, A., Sudore, R., & Schillinger, D. (2011). Interventions to improve patient comprehension in informed consent for medical and surgical procedures: A systematic review. *Medical Decision Making*, *31*(1), 151–173.

Schiepek, G., Tominschek, I., Karch, S., Lutz, J., Mulert, C., Meindl, T., & Pogarell O. (2009). A controlled single case study with repeated fMRI measurements during the treatment of a patient with obsessive-compulsive disorder: Testing the nonlinear dynamics approach to psychotherapy. *World Journal of Biological Psychiatry*, *10*, 658–668.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*(3), 233–242.

Schmidt, M. (2012). Unobtrusive marketing research methods: An overview. *Engineering Management Research*, *1*(2), 172–177.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100.

Schuetter, J., Goel, P., McCorriston, J., Park, J., Senn, M., & Harrower, M. (2013). Autodetection of ancient Arabian tombs in high-resolution satellite imagery. *International Journal of Remote Sensing, 34*(19). Available at www.tandfonline.com/doi/abs/10.1080/01431161.2013.802054#.Uti2IvRDseh

Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly*, *75*(1), 115–124.

Schutz, F. A. B., Je, Y., Richards, C. J., & Choueiri, T. K. (2012). Meta-analysis of randomized controlled trials for the incidence and risk of treatment-related mortality in patients with cancer treated with vascular endothelial growth factor tyrosine kinase inhibitors. *Journal of Clinical Oncology*, *30*, 871–877.

Schwartz, C., & Sprangers, M. A. G. (Eds.) (2000). *Adaptation to changing health: Response shift in quality of life research*. Washington, DC: American Psychological Association.

Schwarz, N., & Hippler, H-J. (2004). Response alternatives: The impact of their choice and presentation order. In P. P. Biemer, R. W. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp.41–56). Hoboken, NJ: John Wiley & Sons.

Science. (2013, October 25) (no authors listed). Science and society: Experts warn against bans on 3D printing, *Science*, *342*, 439.

Seeman J. I., & House M. C. (2010). Influences on authorship issues: An evaluation of receiving, not receiving, and rejecting credit. *Accountability in Research: Policies and Quality Assurance*, *17*, 176–197.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The *Consumer Reports* study. *American Psychologist, 50,* 965–974.

Sen, S. (2012). Socioeconomic status and mental health: What is the causal relationship? *Acta Psychiatrica Scandinavica*, *125*, 187–188.

Sen, S., Duman, R., & Sanacora, G. (2008). Serum brain-derived neurotrophic factor, depression, and antidepressant medications: Meta-analyses and implications. *Biological Psychiatry*, *64*, 527–532.

Serretti, A., & Fabbri, C. (2013). Shared genetics among major psychiatric disorders. *The Lancet, 381*, 1339–1341.

Sesso, H. D., Christen, W. G., Bubes, V., Smith, J. P., MacFadyen, J., Schvartz, M., . . . Gaziano, J. M. (2012). Multivitamins in the prevention of cardiovascular disease in men: The Physicians' Health Study II randomized controlled trial. *Journal of the American Medical Association*, *308*(17), 1751–1760.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *3*, 188–196.

Shanely, R. A., Nieman, D. C., Henson, D. A., Jin, F., Knab, A. M., & Sha, W. (2013). Inflammation and oxidative stress are lower in physically fit and active adults. *Scandinavian Journal of Medicine & Science in Sports*, *23*, 215–223.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78*(3), 298–311.

Shirayama, Y., Andrew, C. H., Chen, S. N., Russell, D. S., & Duman, R. S. (2002). Brain-derived neurotrophic factor produces antidepressant effects in behavioral models of depression. *Journal of Neuroscience*, *2*, 3251–3261.

Shrout, P. E. (Ed.) (1997). Special series: Should significance testing be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*, 1–20.

Siedlinski, M., Boer, J. M. A., Smit, H. A., Postma, D. S., & Boezen, H. M. (2012). Dietary factors and lung function in the general population: wine and resveratrol intake. *European Respiratory Journal*, *39*(2), 385–391.

Sikes, R. S., & Gannon, W. L. (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *Journal of Mammalogy*, *92*(1), 235–253.

Silva, P. A. (1990). The Dunedin Multidisciplinary Health and Development Study: A fifteen year longitudinal study. *Perinatal and Paediatric Epidemiology*, *4*, 76–107.

Silva-Ayçaguer, L. C., Suárez-Gil, P., & Fernández-Somoano, A. (2010). The null hypothesis significance test in health sciences research (1995–2006): Statistical analysis and interpretation. *BMC Medical Research Methodology*, *10*, 44. Available at www.biomedcentral.com/1471-2288/10/44

Simard, V., & Nielsen, T. A. (2005). Sleep paralysis–associated sensed presence as a possible manifestation of social anxiety. *Dreaming*, *15*, 245–260.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy Research*, *22*, 638–647.

Simonsohn, U., Nelson, L., & Simmons, J. (2013). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*. (online posting).

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., . . . Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS One*, *1*(1), e39.

Slavich, G. M., & Cole, S. W. (2013). The emerging field of human social genomics. *Clinical Psychological Science*, *1*, 331–348.

Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N. J., & Whipple, K. (1975). *Psychotherapy versus Behavior Therapy*. Cambridge, MA: Harvard University Press.

Small, G. W., Kepe, V., Siddarth, P., Ercoli, L. M., Merrill, D. A., Donoghue, N., . . . Barrio, J. R. (2013). PET scanning of brain tau in retired National Football League players: Preliminary findings. *The American Journal of Geriatric Psychiatry*, *21*(2), 138–144.

Smet, A. F., & Byrne, R. W. (2013). African elephants can use human pointing cues to find hidden food. *Current Biology*, *23*, 1–5.

Smith, C., & Kanalley, C. (2011, May). Fired Over Facebook: 13 Posts That Got People CANNED. Huffington Post. Available at: www.huffingtonpost.com/2010/07/26/fired-over-facebook-posts_n_659170.html

Smith, E., & Williams-Jones, B. (2012). Authorship and responsibility in health sciences research: A review of procedures for fairly allocating authorship in multi-author studies. *Science and Engineering Ethics*, *18*(2), 199–212.

Smith, G. T., Combs, J. L., & Pearson, C. M. (2012). Brief instruments and short forms. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 395–409). Washington, DC: American Psychological Association.

Smith, J. E., Lawrence, A. D., Diukova, A., Wise, R. G., & Rogers, P. J. (2012). Storm in a coffee cup: caffeine modifies brain activation to social signals of threat. *Social Cognitive and Affective Neuroscience*, *7*(7), 831–840.

Smith, M. L., & Glass G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*, 178–182.

Smith, T., Linnemeyer, R., Scalise, D., & Hamilton, J. (2013). Barriers to outpatient mental health treatment for children and adolescents: Parental perspectives. *Journal of Family Psychotherapy*, *24*(2), 73–92.

Snapinn, S., Chen, M. G., Jiang, Q., & Koutsoukos, T. (2006). Assessment of futility in clinical trials. *Pharmaceutical Statistics*, *5*(4), 273–281.

Solomon, P., Cavanaugh, M. M., & Draine, J. (2009). *Randomized controlled trials: Design and implementation for community-based psychosocial interventions*. New York: Oxford University Press.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*, 137–150.

Solzbacher, S. Böttger, D. Memmesheimer, M. Mussgay, L. & Rüddel, H. (2007). Improving tension regulation in patients with personality disorders, post-traumatic stress disorder and bulimia. In M. J. Sorbi, H. Rüddel, M. E. F. Bühring (Eds.), *Frontiers in stepped eCare: eHealth methods in behavioural and psychosomatic medicine* (pp. 111–119). Utrecht, the Netherlands: University of Utrecht.

Sommers, R., & Miller, F. G. (2013). Foregoing debriefing in deceptive research: Is it ever ethical? *Ethics & Behavior*, *23*(2), 98–116.

Sontag, M. (2012). Research ethics and Institutional Review Boards: The influence of moral constraints on emotion research. *Politics and the Life Sciences*, *31*(1), 67–79.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*(2), 330–348.

Spencer, E. A., Appleby, P. N., Davey, G. K., & Key, T. J. (2002). Validity of self-reported height and weight in 4808 EPIC–Oxford participants. *Public Health Nutrition*, *5*(04), 561–565.

Spiegel, D., Bloom, J. R., Kraemer, H. C., & Gottheil, E. (1989). Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet*, *2* (8668), 888–891.

Spiegel, D., Butler, L. D., Giese-Davis, J., Koopman, C., Miller, E., DiMiceli, S., . . . Kraemer, H. C. (2007). Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer. *Cancer*, *110*(5), 1130–1138.

Spier, R. (2002). The history of the peer-review process. *Trends in Biotechnology, 20,* 357–358.

Spirrison, C. L., & Mauney, L. T. (1994). Acceptability bias: The effects of treatment acceptability on visual analysis of graphed data. *Journal of Psychopathology and Behavioral Assessment, 16*, 85–94.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Sporns, O. (Ed.) (2010). *Analysis and function of large-scale brain networks*. Washington, DC: Society for Neuroscience.

Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, *25*, 225–230.

Stapel, D. A. (2012). *Ontsporing* [Derailment in Dutch]. Amsterdam: Prometheus.

Stead, L. F., Bergson, G., & Lancaster, T. (2008). Physician advice for smoking cessation. *Cochrane Database of Systematic Reviews,* Issue 2 (Art. No.: CD000165).

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267.

Stevenson, C. (2013). A qualitative exploration of relations and interactions between people who are homeless and use drugs and staff in homeless hostel accommodation. *Journal of Substance Use*. Available at http://informahealthcare.com/doi/abs/10.3109/14659891.2012.754508

Stewart, D. W. (2012). Secondary analysis and archival research: Using data collected by others. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 3, pp. 473–484). Washington, DC: American Psychological Association.

Stewart, K. K., Carr, J. E., Brandt, C. W., & McHenry, M. M. (2007). An evaluation of the conservative dual-criterion method for teaching university students to visually inspect AB-design graphs. *Journal of Applied Behavior Analysis, 40,* 713–718.

Stratton, K., Gable, A., Shetty, P., & McCormick, M. (Eds.) (2001). *Immunization safety review: measles-mumps-rubella vaccine and autism*. Washington, DC: National Academies Press.

Strauss, K., Vicari, S., Valeri, G., D'Elia, L., Arima, S., & Fava, L. (2012). Parent inclusion in early intensive behavioral intervention: The influence of parental stress, parent treatment fidelity and parent-mediated generalization of behavior targets on child outcomes. *Research in Developmental Disabilities*, *33*(2), 688–703.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Gosport, Hampshire, UK: Oxford University Press.

Stulz, N., & Lutz, W. (2007). Multidimensional patterns of change in outpatient psychotherapy: The phase model revisited. *Journal of Clinical Psychology, 63,* 817–833.

Su, D., & Li, L. (2011). Trends in the use of complementary and alternative medicine in the United States: 2002–2007. *Journal of Health Care for the Poor and Underserved*, *22*(1), 296–310.

Substance Abuse and Mental Health Services Administration. (2011). SAMHSA announces a working definition of "recovery" from mental disorders and substance abuse disorders. Available at www.samhsa.gov/newsroom/advisories/1112223420.aspx

Sugarman, S. D. (2007). Cases in vaccine court: Legal battles over vaccines and autism. *New England Journal of Medicine*, *357*(13), 1275–1277.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989–1004.

Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E. J., Olsson, A., & Axelsson, J. (2013). Cues of fatigue: Effects of sleep deprivation on facial appearance. *Sleep*, *36*(9), 1355–1360.

Suresh, S. (2011). Moving toward global science. *Science*, *333*, 802.

Swaminathan, H., Horner, R. H., Sugai, G., Smolkowski, L., Hedges, L., & Spaulding, S. A. (2008). Application of generalized least squares regression to measure effect size in single-case research: A technical report. Institute of Education Sciences Technical Report: US Department of Education.

Swift, J. K., & Callahan, J. L. (2009). The impact of client treatment preferences on outcome: A meta-analysis. *Journal of Clinical Psychology*, *65*, 368–381.

Tabibnia, G., Monterosso, J. R., Baicy, K., Aron, A. R., Poldrack, R. A., Chakrapani, S., . . . London, E. D. (2011). Different forms of self-control share a neurocognitive substrate. *Journal of Neuroscience*, *31*(13), 4805–4810.

Takayanagi, Y., Spira, A. P., Roth, K. B., Gallo, J. J., Eaton, W. W., & Mojtabai, R. (2014). Accuracy of reports of lifetime mental and physical disorders: results from the Baltimore Epidemiological Catchment Area study. *JAMA psychiatry, 71*(3), 273–280.

Tang, T. Z., DeRubeis, R. J., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *73*(1), 168–172.

Tanser, F., Bärnighausen, T., Grapsa, E., Zaidi, J., & Newell, M. L. (2013). High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science*, *339*(6122), 966–971.

Taylor, G., McNeill, A., Girling, A., Farley, A., Lindson-Hawley, N., & Aveyard, P. (2014). Change in mental health after smoking cessation: systematic review and meta-analysis. *BMJ*, *348*, g1151.

Taylor, M. F., Marais, I., & Cottman, R. (2012). Patterns of graffiti offending: Towards recognition that graffiti offending is more than 'kids messing around'. *Policing and Society*, *22*(2), 152–168.

Teddlie, C., & Tashakkori, A. (2012). Common "core" characteristics of mixed methods research: A review of critical issues and call for greater convergence. *American Behavioral Scientist*, *56*(6), 774–788.

Telch, M. J., Rosenfield, D., Lee, H. J., & Pai, A. (2012). Emotional reactivity to a single inhalation of 35% carbon dioxide and its association with later symptoms of Posttraumatic Stress Disorder and anxiety in soldiers deployed to Iraq: Reactivity to CO_2 as predictor of PTSD and anxiety. *Archives of General Psychiatry*, *69*(11), 1161–1168.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thomas, L. R., Donovan, D. M., Sigo, R. L., Austin, L., Alan Marlatt, G., & The Suquamish Tribe. (2009). The community pulling together: A tribal community–university partnership project to reduce substance abuse and promote good health in a reservation tribal community. *Journal of Ethnicity in Substance Abuse*, *8*(3), 283–300.

Thomas, S. B., & Quinn, S. C. (1991). The Tuskegee Syphilis Study, 1932 to 1972: Implications for HIV education and AIDS risk education programs in the black community. *American Journal of Public Health*, *81*(11), 1498–1505.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 246–262). Thousand Oaks, CA: Sage

Thompson, J. (2010, July 17). 70s throwback: Lime-yellow fire trucks fade out. *FireRescue1*. [Online article]. Available at www.firerescue1.com/fire-products/fire-apparatus/articles/831990-70s-throwback-Lime-yellow-fire-trucks-fade-out/

Thomson Reuters. (2011). *Journal search: Psychology*. New York: Thomson Reuters.

Thornhill, R., Chapman, J. F., & Gangestad, S. W. (2013). Women's preferences for men's scents associated with testosterone and cortisol levels: Patterns across the ovulatory cycle. *Evolution and Human Behavior*, *34*(3), 216–221.

Thursby, G. (2011). Psychology virtual library: Journals (electronic and print). Retrieved August 2011 from www.vl-site.org/psychology/journals.html

Tingstrom, D., Turner, H., & Wilczynski, S. (2006). The Good Behavior Game: 1969–2002. *Behavior Modification, 30,* 225–253.

Touchette, E., Henegar, A., Godart, N. T., Pryor, L., Falissard, B., Tremblay, R. E., & Côté, S. M. (2011). Subclinical eating disorders and their comorbidity with mood and anxiety disorders in adolescent girls. *Psychiatry Research*, *185*(1), 185–192.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

Triggle, N. (2010, May). MMR doctor struck from register. BBC News. Available at http://news.bbc.co.uk/2/hi/8695267.stm

Trimble, J. E., & Dickson, R. (2005). Ethnic gloss. In C. B. Fisher & Lerner, R. M. (Eds.), *Encyclopedia of applied developmental science* (Vol. 1, pp. 412–415). Thousand Oaks, CA: Sage.

Tsang, R., Colley, L., & Lynd, L. D. (2009). Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of Clinical Epidemiology*, *62*, 609–616.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

Turke-Browne, B. (2013). Functional interactions as big data in the human brain. *Science*, *342*, 580–584.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*(3), 252–260.

Turner, M. G., Exum, M. L., Brame, R., & Holt, T. J. (2013). Bullying victimization and adolescent mental health: General and typological effects across sex. *Journal of Criminal Justice*, *41*(1), 53–59.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies. How to deal with missing data. *Journal of Clinical Epidemiology, 55*(4), 329–337.

Twohig, M. P., Shoenberger, D., & Hayes, S. C. (2007). A preliminary investigation of acceptance and commitment therapy as a treatment for marijuana dependence in adults. *Journal of Applied Behavior Analysis, 40,* 619–632.

Underwood, E. (2013). New tools light up the intricacies of the brain. *Science*, *342*, 917–918.

United States Census Bureau. (2012, December). U.S. Census Bureau projections show a slower growing, older, more diverse nation a half century from now. Washington, DC: Department of Commerce. Available at www.census.gov/newsroom/releases/archives/population/cb12-243.html

United States Department of Defense. (2012, April). Sexual assault prevention and response. *Annual Report on Sexual Assault in the Military*. Washington, DC. Department of Defense.

United States Department of Health and Human Services. (1979). *The Belmont report*. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Available at www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

United States Department of Health and Human Services. (1983). National Institutes of Health, Office for Protection from Research Risks. *Code of federal regulations: Part 46: Protection of human subjects*. Washington, DC: US Government Printing Office.

United States Department of Health and Human Services. (2009a). Code of Federal regulations: Part 46, Protection of human subjects (45 CFR 45.116). Available at www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.116

United States Department of Health and Human Resources. (2009b). Federal Coordinating Council for Comparative Effectiveness

Research. Report to the President and the Congress. Available at www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf

United States Department of Health and Human Services. (2012). Office of Research Integrity. Available at http://ori.hhs.gov/about-ori

University of Virginia. (2013). Institutional Review Board for Social and Behavioral Sciences. Sample debriefing statement. Available at www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html

Uziel, L. (2010). Rethinking social desirability scales from impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*, 468–472.

Valentine, J. C. (2012). Meta-analysis. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 3, pp. 485–499). Washington, DC: American Psychological Association.

Van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., & Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: Evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychological Medicine*, *39*(2), 179–195.

Vartanian, L. R., Schwartz, M. B., & Brownell, K. D. (2007). Effects of soft drink consumption on nutrition and health: A systematic review and meta-analysis. *American Journal of Public Health*, *97*(4), 667–675.

Vassos, E., Collier, D. A., & Fazel, S. (2014). Systematic meta-analyses and field synopsis of genetic association studies of violence and aggression. *Molecular Psychiatry*, *19*(4), 471–477.

Vaz, L. M., Eng, E., Maman, S., Tshikandu, T., & Behets, F. (2010). Telling children they have HIV: Lessons learned from findings of a qualitative study in sub-Saharan Africa. *AIDS Patient Care and STDs*, *24*(4), 247–256.

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly*, *37*(1), 21–54.

Ventimiglia, M., & MacDonald, D. A. (2012). An examination of the factorial dimensionality of the Marlowe Crowne social desirability scale. *Personality and Individual Differences*, *52*(4), 487–491.

Verdonk, P., Beaufils, P., Bellemans, J., Djian, P., Heinrichs, E. L., Huysse, W., . . . Pðssler, H. (2012). Successful treatment of painful irreparable partial meniscal defects with a polyurethane scaffold two-year safety and clinical outcomes. *The American Journal of Sports Medicine*, *40*(4), 844–853.

Vieyra, M., Strickland, D., & Timmerman, B. (2013). Patterns in plagiarism and patchwriting in science and engineering graduate students' research proposals. *International Journal for Educational Integrity*, *9*(1), 35–49.

Vøllestad, J., Sivertsen, B., & Nielsen, G. H. (2011). Mindfulness-based stress reduction for patients with anxiety disorders: Evaluation in a randomized controlled trial. *Behaviour Research and Therapy*, *49(4)*, 281–288.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., . . . Walker-Smith, J. A. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637–641.

Walker, C. K., Anderson, K. W., Milano, K. M., Ye, S., Tancredi, D. J., Pessah, I. N., . . . Kliman, H. J. (2013). Trophoblast inclusions are significantly increased in the placentas of children in families at risk for autism. *Biological Psychiatry*, *74*(3), 204–211.

Walker, E. F., Trotman, H. D., Pearce, B. D., Addington, J., Cadenhead, K. S., Cornblatt, B. A., . . . Woods, S. W. (2013). Cortisol levels and risk for psychosis: Initial findings from the North American Prodrome Longitudinal Study. *Biological Psychiatry*, *74*(8), 410–417.

Wallace, M. L., Frank, E. J., & Kraemer, H. C. (2013). A novel approach for developing and interpreting moderator profiles in randomized controlled trials. *JAMA-Psychiatry*, *70*(11), 1241–1247.

Wallerstein, R. S. (1986). *Forty-two lives in treatment: A study of psychoanalysis and psychotherapy*. New York: Guilford.

Walsh, R. (2011). Lifestyle and mental health. *American Psychologist*, *66*, 579–592.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011, July). "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security.* Article No. 10. New York: Association for Computing Machinery.

Ward, A. C. (2009). The role of causal criteria in causal inferences: Bradford Hill's "aspects of association." *Epidemiologic Perspectives & Innovations*, *6*, 2. Available at http://archive.biomedcentral.com/1742-5573/content/6/1/2

Warran, C. (2011). 10 people who lost jobs over social media mistake. *Mashable*. Available at http://mashable.com/2011/06/16/weinergate-social-media-job-loss/#_

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology: Volume 10: Assessment psychology* (2nd ed., pp. 50–81). New York: John Wiley & Sons.

Watanabe, M., & Aoki, M. (2014, January 10). Researcher: Test data falsified in major Alzheimer's disease project, *The Ashai Shimbun* Available at https://ajw.asahi.com/article/behind_news/social_affairs/AJ201401100085

Watson, D. (2012). Objective tests as instruments of psychological theory and research. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, pp. 349–369). Washington, DC: American Psychological Association.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, *3*, 1–14.

Watson, T. S., Meeks, C., Dufrene, B., & Lindsay, C. (2002). Sibling thumb sucking: Effects of treatment for targeted and untargeted siblings. *Behavior Modification*, *26*, 412–423.

Web of Knowledge. (2011). *2010 Journal citation reports*. New York: Thomson Reuters. Retrieved August 2011 from http://wokinfo.com/products_tools/analytical/jcr/

Web of Science. (2011). *2010 Journal citation reports*. New York: Thomson Reuters. Available at http://wokinfo.com/products_tools/analytical/jcr/

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (revised edition). Thousand Oaks, CA: Sage Publications

Webb, J. R., Schroeder, M. I., Chee, C., Dial, D., Hana, R., Jefee, H., . . . Molitor, P. (2013). Left-handedness among a community sample of psychiatric outpatients suffering from mood and psychotic disorders. *SAGE Open*, *3*(4), 2158244013503166.

Wechsler, M. E., Kelley, J. M., Boyd, I. O. E., Dutile, S., Marigowda, G., M. B., Kirsch, I., . . . Kaptchuk, T. J. (2011). Active albuterol or placebo, sham acupuncture, or no intervention in asthma. *New England Journal of Medicine*, *365*, 119–126.

Wehby, J. H., & Hollahan, M. S. (2000). Effects of high-probability requests on latency to initiate academic tasks. *Journal of Applied Behavior Analysis*, *33*, 259–262.

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. *JAMA Psychiatry, 70*(7), 750–761.

Weisz, J. R., & Kazdin, A. E. (Eds.) (2010). *Evidence-based psychotherapies for children and adolescents* (2nd ed.). New York: Guilford Press.

Weisz, J. R., & Kazdin, A. E. (Eds.) (2010). *Evidence-based psychotherapies for children and adolescents* (2nd ed.). New York: Guilford Press.

Weisz, J. R., Ng, M. Y., & Bearman, S. K. (2014). Odd couple? Re-envisioning the relation between science and practice in the dissemination and implementation era. *Clinical Psychological Science*, *2*, 58–74.

Welfel, E. R. (2013). *Ethics in counseling and psychotherapy: Standards, research, and emerging issues* (5th ed.). Belmont, CA: Brooks Cole/ Cengage Learning.

Welham, J., Isohanni, M., Jones, P., & McGrath, J. (2009). The antecedents of schizophrenia: A review of birth cohort studies. *Schizophrenia Bulletin*, *35*(3), 603–623.

Wells, G. L., & Loftus, E. F. (2013). Eyewitness testimony memory for people and events. In I.B. Weiner (Ed.), *Handbook of research. Volume 11. Forensic psychology* (pp. 617–629). New York: John Wiley & Sons.

Wells, G. L., & Penrod, S. S. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfeld & S.D. Penrod (Eds.), *Research methods in forensic psychology* (pp 237–256). Hoboken, NJ: John Wiley & Sons.

Wells, J. E., & Horwood, L. J. (2004). How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports. *Psychological Medicine*, *34*, 1001–1011.

Werner, E. E., & Smith, R. S. (1982). *Vulnerable, but invincible: A longitudinal study of resilient children and youth*. New York: McGraw-Hill.

Wertz, F. J., Charmaz, K., & McMullen, L. M. (2011). *Five ways of doing qualitative analysis: Phenomenological psychology, grounded theory, discourse analysis, narrative research, and intuitive inquiry*. New York: Guilford Press.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*, 267–270.

Westermeyer, J., & Canive, J. (2012). Posttraumatic stress disorder and its comorbidities among American Indian veterans. *Community Mental Health Journal*, *49*, 704–708.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*(3), 291–298.

Whalen, C., Schreibman, L., & Ingersoll, B. (2006). The collateral effects of joint attention training on social initiations, positive affect, imitation, and spontaneous speech for young children with Autism. *Journal of Autism and Developmental Disorders, 36*, 655–664.

Whitley, R., & Drake, R. (2010). Recovery: A dimensional approach. *Psychiatric Services*, *61*(12), 1248–1250.

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, *11*(4), 522–537.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE 6*(11): e26828.

Wichstrøm, L., Berg-Nielsen, T. S., Angold, A., Egger, H. L., Solheim, E., & Sveen, T. H. (2012). Prevalence of psychiatric disorders in preschoolers. *Journal of Child Psychology and Psychiatry, 53*, 695–705.

Widom, C. S., & Shepard, R. L. (1996). Accuracy of adult recollections of childhood victimization: Part 1. Childhood physical abuse. *Psychological Assessment*, *8*, 412–421.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wilkinson, M. (2013). Testing the null hypothesis: The forgotten legacy of Karl Popper? *Journal of Sports Sciences*, *31*(9), 919–920.

Williams, M. T., Domanico, J., Marques, L., Leblanc, N. J., & Turkheimer, E. (2012). Barriers to treatment among African Americans with obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *26*(4), 555–563.

Wilson, A. M., Lowe, J. C., Roskilly, K., Hudson, P. E., Golabek, K. A., & McNutt, J. W. (2013). Locomotion dynamics of hunting in wild cheetahs. *Nature, 498*, 185–189.

Wilson, F. A., & Stimpson, J. P. (2010). Trends in fatalities from distracted driving in the United States, 1999 to 2008. A*merican Journal of Public Health*, *100*(11), 2213–2219.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd. ed.). New York: McGraw-Hill.

Wipfli, B., Landers, D., Nagoshi, C., & Ringenbach, S. (2011). An examination of serotonin and psychological variables in the relationship between exercise and mental health. *Scandinavian Journal of Medicine & Science in Sports, 21*, 474–481.

Wiser, M. J., Ribeck, N., & Lenski, R. F. (2013, November). Long-term dynamics of adaptation in asexual populations. *Science*. Available at www.sciencemag.org/content/early/2013/11/18/science.1243357/abstract

Wislar, J. S., Flanagin, A., Fontanarosa, P. B., & DeAngelis, C. D. (2011). Honorary and ghost authorship in high impact biomedical journals: A cross sectional survey. *British Medical Journal*, *343*. d6128.

Wittenbrink, B., & Schwarz, N. (Eds.) (2007). *Implicit measures of attitudes*. New York: Guilford Press.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of g-theory methods for modeling multitrait–multimethod data clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134–161.

Wolf, M. M. (1978). Social validity: The case of subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.

Wolf, S. M., Annas, G. J., & Elias, S. (2013). Patient autonomy and incidental findings in clinical genomics. *Science*, *340*, 1049–1050.

Wolff, E., Gaudlitz, K., von Lindenberger, B. L., Plag, J., Heinz, A., & Ströhle, A. (2011). Exercise and physical activity in mental disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *261*(2), 186–191.

Wonkam, A., Fieggen, K., & Ramesar, R. (2010). Beyond the Caster Semenya controversy: The case of the use of genetics for gender testing in sport. *Journal of Genetic Counseling*, *19*(6), 545–548.

Woolcott, J. C., Richardson, K. J., Wiens, M. O., Patel, B., Marin, J., Khan, K. M., & Marra, C. A. (2009). Meta-analysis of the impact of 9 medication classes on falls in elderly persons. *Archives of Internal Medicine*, *169*(21), 1952–1960.

Woolf, S. W. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association*, *299*(2), 211–213.

World Bank. (1993). *World development report 1993: Investing in health*. New York: Oxford University Press.

World Federation for Mental Health. (2011). *The great push: Investing in mental health.* World Federation for Mental Health. World Health Organization (2008). *Task shifting: Global recommendations and guidelines.* Geneva: WHO.

World Health Organization. (2008). *The global burden of disease: 2004 update.* Geneva: WHO.

World Health Organization. (2010). *International classification of diseases-10* (4th ed.). Geneva: WHO.

World Health Organization. (2013). Metrics: disability-adjusted life year (DALY): Quantifying the Burden of disease from mortality and morbidity. Available at www.who.int/healthinfo/global_burden_disease/metrics_daly/en/

World Medical Association. (2013). WMA Declaration of Helsinki: Ethical principles for medical research involving human subjects. Available at www.wma.net/en/30publications/10policies/b3/

Wrzus, C., Hänel, M., Wagner, J., & Neyer, F. J. (2013). Social network changes and life events across the life span: A meta-analysis. *Psychological Bulletin*, *139*(1), 53–80.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, *18*(7), 600–606.

Xu, J., & Roberts, R. E. (2010). The power of positive emotions: It's a matter of life or death—Subjective well-being and longevity over 28 years in a general population. *Health Psychology*, *29*, 9–19.

Yamada, T., Hara, K., Umematsu, H., & Kadowaki, T. (2013). Male pattern baldness and its association with coronary heart disease: A meta-analysis. *British Medical Journal*, *3*(4). Available at http://bmjopen.bmj.com/content/3/4/e002537.short

Ybarra, G. J., Passman, R. H., & Eisenberg, C. S. (2000). The presence of security blankets or mothers (or both) affects distress during pediatric examinations. *Journal of Consulting and Clinical Psychology*, *68*, 322–330.

Yehuda, R., & Seckl, J. (2011). Minireview: stress-related psychiatric disorders with low cortisol levels: A metabolic hypothesis. *Endocrinology*, *152*(12), 4496–4503.

Yong, E. (2012a). Bad copy: In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature*, *485*, 298–300.

Yong, E. (2012b). Nobel laureate challenges psychologists to clean up their act. *Nature News,* Nature Publishing Group.

Zalsman, G. I. L., Patya, M., Frisch, A., Ofek, H., Schapir, L., Blum, I., . . . Tyano, S. (2011). Association of polymorphisms of the serotonergic pathways with clinical traits of impulsive-aggression and suicidality in adolescents: A multi-center study. *World Journal of Biological Psychiatry*, *12*(1), 33–41.

Zamani-Alavijeh, F., Bazargan, M., Shafiei, A., & Bazargan-Hejazi, S. (2011). The frequency and predictors of helmet use among Iranian motorcyclists: A quantitative and qualitative study. *Accident Analysis & Prevention*, *43*(4), 1562–1569.

Zedeck, S. (2014). *APA Dictionary of statistics and research methods.* Washington, DC: American Psychological Association.

Zemore, S. E. (2012). The effect of social desirability on reported motivation, substance use severity, and treatment attendance. *Journal of Substance Abuse Treatment*, *42*(4), 400–412.

Zilliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.

Zimbardo, P. G. (2007). *The Lucifer effect: Understanding how good people turn evil*. New York: Random House.

Zimbardo, P. G., & Cross, A. B. (1971). *Stanford prison experiment*. Stanford University. Available at www-sul.stanford.edu

Ziv, M., Tomer, R., Defrin, R., & Hendler, T. (2010). Individual sensitivity to pain expectancy is related to differential activation of the hippocampus and amygdala. *Human Brain Mapping*, *31*(2), 326–338.

End Notes

Chapter 1

1. In keeping with guidelines for publication of psychological research, the terms "participants" and "subjects" will be used interchangeably to refer to those individuals who serve in the study (American Psychological Association [APA], 2010b). Throughout the text, the terms "subjects," "participants," and "clients" will be used to delineate those persons who are being studied, i.e., those who participate in research and provide the data. Participants in research can include investigators (who design the study) and experimenters (who administer the conditions) and, in an important sense, consumers of research (other professionals, the public at large).
2. Puerperal fever is a form of septicemia also referred to as sepsis, which is a serious medical condition usually bacterial in nature that can lead to death. The cause is an extensive immune response to the infection. Many chemicals released to fight the infection cause the widespread inflammation. In turn, this can result in organ damage and blood clotting that reduces blood flow to other organs and the limbs. As the infection progresses, there can be a severe drop in blood pressure (referred to as septic shock) and from that failure of major organs (liver, kidneys, lungs) and death. Sepsis occurs in 1% to 2% of all hospitalizations in the United States and affects approximately 750,000 people each year (see www.webmd.com/a-to-z-guides/sepsis-septicemia-blood-infection). The disorder can be treated, but an intensive set of interventions may be needed both to get rid of infection and to manage the individual symptom problems and organ failure that emerges.
3. The outcome of all of this during Semmelweis's lifetime was tragic for him. Among the ensuing events, he was tricked into being hospitalized, was beaten in the hospital, and within 2 weeks of the beating died (at the age of 47). The tragic irony—he died as a result of infections from the beating (of sepsis). But in later years, his contributions were recognized as extraordinary; his image is on scores of coins and postage stamps; and there are statues, biographies, a musical score, museums, gynecological clinics, and a medical school named after him. Moreover, he is known as "the savior of mothers." And in medical training, the story is told to warn against the dangers of arrogance. Of course, not arrogance on the part of Semmelweis but on the part of current science and medical practice at the time in which his discovery was dismissed. Occasionally, the term "Semmelweis Effect" or "Semmelweis Reflex" is used to mean that some new idea is automatically or quickly dismissed because it goes against current views, beliefs, and paradigms.

Chapter 2

1. The formal delineation of various types of experimental validity owes its origin to Donald Campbell (1916–1996), a psychologist who contributed enormously to methodology (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979). He referred to one type of validity as "statistical conclusion" validity, which has to do with problems that emerge in a study related to data evaluation. These problems can interfere with drawing clear conclusions. I have substituted the term "data-evaluation validity" in place of Campbell's "statistical conclusion validity" for three reasons: First, not all data problems involve statistical tests. Second, many experimental designs (some of which we cover) do not use statistical tests at all in drawing inferences (Kazdin, 2011). Yet, these studies can still have problems in interpreting the results based on how the data are evaluated. Finally, I believe the term is clearer—what kind of problems emerge in a study? The answer is nicely covered by the more straightforward term "data evaluation" than "statistical conclusion."
2. Throughout the course, references will made to studies and investigations. I use these as generic terms to refer to any scientific research or empirical study. More specific terms will be used occasionally. Experiment or *true experiment* is a term that refers to a study in which the investigator can assign participants randomly to conditions and manipulates a variable of interest (e.g., some experience provided to the participant, some intervention). *Observational study* is a term used to reflect an investigation in which one selects groups or conditions (e.g., depressed vs. nondepressed patients) or studies a phenomenon over time (e.g., impact of early trauma) where one does not directly manipulate the variable of interest but observes manipulations "by nature" or conditions not controlled by the investigator. More will be said about true-experiments and observational studies and the distinctions.
3. Most of the prior threats refer to conditions that apply to all groups within the study (e.g., history, maturation) and that could explain the pattern of findings. Yet, history, maturation, and the other threats also may affect groups differently in a given study. Whenever threats to internal validity vary for the different groups within the study, i.e., apply to only one of the groups, these are referred to as *combinations of selection and that other threat*. Another way to refer to this is to say that the threat interacts with (differentially applies to) groups (e.g., experimental and control conditions). An example selection *x* history would mean that one of the groups has an historical experience (exposure to some event in or outside of the investigation) that the other group did not have and that experience might plausibly explain the results. The threat is referred to as selection *x* history because the threat (history) was selective and applies to only one (or some but not all) of the groups. As a general statement, if any single influence (e.g., history, testing) applies to only one of the groups or applies in different ways to the groups, the threat involves a combination of selection and that other threat. Some practices we regard as routine such as carefully supervising implementation of an intervention and ensuring subjects are treated in identical ways except for the experimental manipulation are conducted in part to control for selection *x* history threats. I mention this type of threat only in passing because it is as common as the other threats. Yet mention of this threat is important to mention to clarify other threats we have already discussed. History, maturation, and other threats we discussed before refer to those situations in which changes in all of the groups can be explained by the threat. Less likely but possible is that there is that one of these threats somehow applies to only one group. To even pose this possibility, there would need to be a stark event or clear variation in how the groups were treated apart from the experimental manipulation.
4. The groupings routinely used to characterize ethnic and cultural groups can be readily challenged. For example, in the United States, one speaks of Hispanic American or Latino American participants. This is a highly heterogeneous group, and it makes unclear sense based on genetics, culture, and country of origin to form this larger class group. The term "ethnic gloss" is used to refer to the overgeneralization and oversimplification of these generic terms (Trimble & Dicksen, 2005). That is, current groupings in research that describe participants (European American, African American, Asian American, European American, Hispanic American) *gloss* over meaningful subgroupings.

5. In discussion of treatment and prevention, including psychology, psychiatry, and medicine, a distinction is often made between efficacy and effectiveness (e.g., Eichler et al., 2011). *Efficacy* refers to treatment outcomes obtained in controlled psychotherapy studies that are conducted under laboratory and quasi-laboratory conditions (e.g., cases are recruited who are homogeneous and who may show a narrow range of problems, treatment is specified in manual form, and treatment delivery is closely supervised and monitored). *Effectiveness* refers to treatment outcomes obtained in clinical settings where the usual control procedures are not implemented. I use the term "effectiveness" here generically to mean having impact on the problem that has been treated.
6. In the original source here, the authors do not make the distinction between testing and pretest sensitization, and these concepts in that source are not optimally clear.
7. Pretest sensitization is the primary external validity concern here. There is something to be aware of as a more esoteric variation. Even when a pretest is not used, it is possible that assessment may influence the results. The posttest might sensitize subjects to the previous intervention that they have received and yield results that would not have been evident without the assessment. This effect, referred to as *posttest sensitization* (Bracht & Glass, 1968), is very similar to pretest sensitization where test administration may crystallize a particular reaction on the part of the subject. With posttest sensitization, assessment constitutes a necessary condition for the experimental manipulation to show its effect. The effects might be latent, minimal, or not appear at all if the subjects did not know they were being assessed on some posttest. Subjects can tell they are being assessed (obtrusive assessment) and prompted to think of violence in a way they might not have otherwise thought (sensitized). As a threat to external validity, posttest sensitization raises the question of whether the results would extend to measures that subjects could not associate with intervention or measures that were completely out of their awareness. The effect of posttest sensitization is slightly more difficult to assess and control than is pretest sensitization because it requires the use of unobtrusive measures of treatment effects and a comparison of the results across measures varying in whether or not or degree of their obtrusiveness.

Chapter 3

1. Construct validity is a more familiar term in the context of test development and validation (e.g., Grimm & Widaman, 2012). An investigator may develop a psychological test to measure anxiety. Several types of studies are completed to establish the construct validity, that it is anxiety that the scale measures, rather than some other construct (e.g., intelligence, deviance, socially desirable responding, honesty, altruism). Thus, in the use of test development, construct validity refers to the explanation of the measure or the dimension that it assesses. In a parallel way, construct validity of an experiment refers to the explanation of the outcome.
2. Sometimes the concept of a triple-blind study is used in which case the subjects, doctors who administer the drugs, and those who oversee the study (e.g., experimenters who monitor drug administration, or investigators responsible for the study) do not know precisely which subjects receive the medication. Which subjects are to receive the medication or placebo is coded (e.g., by patient number, "drug" number for that day), and these codes are stored. When all the results are in and no bias can be conveyed through interactions with the subjects or the doctors, the codes are revealed so that the data can be analyzed. (I think the reader would want to know among the innovations of my dissertation was the use of a quadruple-blind procedure where no one—and I mean no one—knew what condition anyone received. This meant that I could never decide who was in the experimental or control group and hence could not make statistical comparisons. I liked the innovation because it eliminated all sorts of potential biases, maybe even some that have not been invented yet. My dissertation committee thought the inability to report any results was too large a price to pay.)
3. For purposes of discussion, we shall consider the *investigator* as the person who has the responsibility for planning and designing the study and the *experimenter* as the person who is actively running the subjects and carrying out the procedures. This distinction is helpful despite the fact that the investigator and experimenter are occasionally the same person and that multiple persons in a project may vary in the extent to which they share these roles. We focus here on the experimenter to emphasize the person in direct contact with the subjects.
4. Statistical conclusion validity is a term the reader also ought to know and use as a category for classifying the threats related to quantitative data evaluation (Cook & Campbell, 1979). The term is reasonable because the vast majority of psychological research uses statistics to make inferences about the impact of an experimental condition or manipulation. A preferable term, from my perspective, is *data-evaluation validity* because not all data evaluation in psychology is based on inferential statistics. Two examples discussed later include qualitative research and single-case experimental designs, methodologies that depart from the usual group research used in psychology. In these methodologies, statistical evaluation can be, but usually is not, used to draw inferences from the data.
5. Effect size as delineated here is one of many ways of estimating the magnitude or strength of the relationship. The version is among the most familiar and is referred to as Cohen's *d*. The measure illustrates well the importance of effect size and how methodological practices (e.g., sloppy ones) can translate to statistical issues and weaken the likely results that will be found.
6. Science teaches humility because one does not always know what is and what is not related and the unexpected ought to be expected. Shoe size seems silly in this example, and of course, there is no need to assess and screen cases based on shoe size unless one has a hypothesis or evidence points strongly to the prospect that it might make a difference. So I stand by the shoe size example. Yet, shoe size could relate to something that is important in a study. For example, reading ability correlates with shoe size (National Research Council, 2001). One might make a career out of identifying why, but a possible explanation is that children with smaller shoe sizes read less well and that may be influenced by health, socioeconomic disadvantage, and poor nutrition. These latter influences also are likely to relate to exposure to books and reading opportunities. Another explanation would be that shoe size is strongly controlled genetically and might be associated with other physical and psychological attributes that share some common link.

Chapter 4

1. There is an organization called Workaholic's Anonymous (after Alcoholics' Anonymous). The Web site has a questionnaire to identify whether one is a workaholic (www.workaholics-anonymous.org/page.php?page=knowing). This is an excellent example to cite for a course on research methods because it is not clear that the measure to identify oneself as a workaholic on the Web site has any reliability or validity data and would meet minimal requirements for a psychological measure, although there are empirically developed and evaluated measures of the concept (e.g., Aziz, Uhrich, Wuensch, & Swords, 2013; Schaufeli, Shimazu, & Taris, 2009). And in keeping with the topic of this chapter, one research idea would be to see if there is a group of individuals who can be identified as "workaholics" and what other (nonwork) characteristics they have that would be interesting. Another line of work would be to have people rate various vignettes (e.g., pictures of individuals on a computer screen with overlaid descriptions of them and their behavior). One could experimentally manipulate the descriptions to identify what cause people to label others as workaholics. There are many

other groups modeled after Alcoholics Anonymous, at least in organizational names. They all being with the name of the issue (e.g., Proscrastinators, Emotions, Anonymous, Marijuana, Sexaholics, Overeaters) followed by Anonymous. (I tried to start a group named "Methodologists Anonymous," but everyone sent in their real name and thus missed the point.)

2. Investigations of assessment devices can appear in many journals within psychology. However, some journals focus exclusively or almost exclusively on measures and their investigation. Prominent examples include *Psychological Assessment*, *Journal of Personality Assessment*, and *Behavioral Assessment*.

3. Epidemiology refers to the study and the distribution of diseases and related conditions and the factors that influence the distribution. Research focuses on associations between characteristics and diseases and the nature of these associations (e.g., risk factors, cause). The study of clinical disorders from an epidemiological perspective, an area sometimes referred to as psychiatric epidemiology, is directly relevant to many topics of interest in clinical psychology.

4. Arguably the most influential source of criteria to infer cause in science was provided by Sir Austin Bradford Hill (1897–1991), a pioneer in medical statistics and epidemiology. His classic paper (Hill, 1965) identified nine criteria for inferring cause: (1) Strength of association, (2) Consistency, (3) Specificity, (4) Temporality, (5) Biological gradient, (6) Plausibility, (7) Coherence, (8) Experiment, and (9) Analogy (causal relation demonstrated on a closely related topic). (These are enumerated in Table 4.3.) These are referred to as the "Bradford Hill criteria" and have exerted enormous influence on scientific research and continue to be the basis for many articles and texts discussing causality and how that is inferred in scientific research (e.g., see Höfler, 2005, for an excellent presentation and summary).

5. Brain-derived neurotrophic factor (BDNF) is a protein dispersed and secreted throughout the body. As a neurotrophin, BDNF stimulates growth and differentiation of new neurons and synapses (neurogenesis) in the brain and is active especially in the hippocampus, cortex, and basal forebrain, and other areas central to learning, memory, and higher order thinking (see Duman & Aghajanian, 2012).

6. Polymorphism in the context discussed here refers to variation in a specific genetic characteristic, in this case related to a particular neurotransmitter receptor. A more familiar polymorphism is blood type (e.g., ABO grouping). Polymorphisms are an active area of research because they serve as moderators for all sorts of critical processes (e.g., immune system, metabolism of drugs, and susceptibility to eating disorders).

7. A concern about much of clinically relevant research is that it is not being translated and that findings are slow to reach bedside or community. A term that has been coined to denote this is "bench to bookshelf," which refers to the fact that research often goes from the lab to publication in a journal (Insel, 2013). That is the critical criterion in relation to success in academia and may complete with translational goals.

8. Translational research is of enormous interest, and this is reflected now in scores of professional organizations, such as the Association for Translational Science (www.ctssociety.org/) and the Society for Clinical and Translational Science (http://community.sciencecareers.org/ctscinet/partners/scts/), just to mention two examples. Moreover, there are now many professional journals on the topic such as *Translational Psychiatry*, *Translational Research, Science Translational Medicine,* and *American Journal of Translational Research*, and these too are just a few examples. Another area of work closely related to translational research is called Implementation Science (www.fic.nih.gov/News/Events/implementation-science/Pages/faqs.aspx). Implementation science focuses on the movement from evidence-based programs to application (e.g., how to do that, what implementation strategies are effective, how to adapt findings from controlled research to "real" world settings). Sometimes this is characterized as "research to programs" and "research to policy." Although the topic is beyond the goals of the present chapter, the delineation of implementation science conveys the attention and concern in moving findings from research to application.

9. *Caenorhabditis elegans* is a roundworm (nematode) that has many organ systems similar to those of other animals; they have been used as an animal model for a research on a variety of topics (e.g., genetics, aging, learning, and memory) (see database on *C. elegans* research at http://en.wikipedia.org/wiki/WormBase).

Chapter 5

1. Epidemiology is a scientific discipline that focuses on factors and conditions that influence the frequency and distribution of disease, injury, and other health-related events and their causes within the population. Topics and methods of epidemiology overlap with clinical psychology. For example, psychiatric epidemiology is a specialty area that studies the distribution and factors associated with psychological dysfunction in the population. Also, many research designs (e.g., observational designs), well developed in epidemiology are used in clinical psychology, usually on a smaller scale and with less interest in characterizing populations.

2. Occasionally I make reference to a clinic in which I work. This is the Yale Parenting Center, a clinical service for children and families. The Center serves two broad populations of children ages 2–15. The first group consists of children who are referred clinically for oppositional, aggressive, and antisocial behavior. Two evidence-based treatments (variations of cognitive problem-solving skills training and parent management training) are provided (see Kazdin, 2010). The second group of children are not experiencing clinical dysfunction; rather their parents seek help with the normal challenges of parenting (e.g., toilet training, doing homework, teen "attitude") (see Kazdin & Rotella, 2008, 2013). Because many clinical dysfunctions are on a continuum with normative behavior, the populations are easily distinguished at the margins but occasionally blend.

3. There are scores of websites that can provide random numbers to be used for research (e.g., www.Random.org; www.randomizer.org/form.htm; www.psychicscience.org/random.aspx). I am almost certain that the reader is eager to learn that many—probably most or even almost all—random numbers tables or computer-generated sequences of numbers are not truly random. (These are the topics often discussed at late-night methodology parties [that usually end by 8:00 pm].) For example, computers follow algorithms and any particular algorithm influences or rather dictates the next number in the sequence. Technically this is not purely random. An exception is www.Random.org, which generates numbers in the following way. A radio is tuned to an unused frequency. Static is generated from the atmosphere on that station (as one hears static on a radio). The fluctuating static is converted into numbers that are unpredictable and random (from one to the next). These intricacies are not needed for research, but it is important to know that what is called random often is not. Also, the goal is to make implausible the likelihood of selection bias in constructing groups. Virtually all random numbers tables will do that but making implausible does not mean groups will be perfectly equivalent. Indeed, true randomness guarantees that occasionally they will not be.

4. Propensity score matching has many options and methods that are beyond the scope of the present chapter. There are excellent introductions to the analyses, including both special journal articles (e.g., Lane, To, Shelley, & Henson, 2013; Steiner, Cook, Shadish, & Clark, 2010), special issues of journals (e.g., Austin, 2011), and books (e.g., Holmes, 2014; Rosenbaum, 2010). In addition, propensity analyses are available in many commonly used statistical software packages.

5. In most psychology studies where RCTs are used, subjects are assigned randomly to conditions as discussed here. Increasingly research is being conducted on a larger scale and in naturalistic settings. In such cases, many different sites or clinics are used. The setting (e.g., clinic, village, hospital, or other large unit) becomes the focus of the assignment. The design is called a *cluster randomized controlled trial*. The clusters (or settings) are assigned randomly. For example, a large-scale RCT of treatment for anxiety and depression in India assigned 24 public and private clinical services to a special stepped care intervention administered by lay counselors or treatment as usual (Patel et al., 2010) and this was a cluster RCT where the settings and not the individual patients (>2,700 participants) were assigned. All patients in a given setting received the condition to which the setting was assigned. The cluster randomized controlled design has been used heavily in the context of evaluating treatments for HIV/AIDs in developing countries but has been extended well beyond that focus (Osrin et al., 2009).
6. Although the main effects of treatment, order, and groups can be extracted from Latin Square analyses, interactions among these effects present special problems that are beyond the scope of the present chapter. For a discussion of procedures to select or to form Latin Squares for a given experiment and for a table of various squares, the interested reader is referred to other sources (Fisher & Yates, 1963; Kirk, 1994). (For a discussion of strategies for data analyses, there are many excellent and useful resources available, including the seminal paper on the topic [Grant, 1948], a classic text [e.g., Winer, Brown, & Michels, 1991], and resources on the Web [e.g., www.itl.nist.gov/div898/handbook/pri/section3/pri3321.htm; http://statpages.org/latinsq.html]. Also commonly used statistical software packages have options for use of Latin Squares.)

Chapter 6

1. Personalized medicine has as its goal individualizing treatment based on characteristics of each patient. This suggests that the profile of each individual (e.g., based on their diverse biological and other characteristics) will influence the treatment decision. This is a goal. A step toward that goal might be aptly characterized as "moderated medicine," rather than personalized medicine. The difference is that a moderator is not at the level of individuals but of subgroups, i.e., individuals who share a given characteristic. A moderator is identified that influences the effectiveness of treatment and that moderator is used to make decisions. For example, attention has been particularly great in cancer treatment where the goal is to identify genetic or other biological characteristics ("biomarkers" as they are called) that influence responsiveness to treatment. And one or two such biomarkers have been identified and used (and are moderators). Individuals with a given biomarker or two fall into a subgroup that might profit from treatment; those without the biomarker may not. The difficulty is that there are scores of biomarkers and profiling individuals on all of them and making highly individualized decisions is a more complex task than focusing on one or two markers (see Roukos, 2009). It is likely that research will move from using one moderator (one biomarker or psychological characteristic), to a few moderators, and then hopefully to multiple moderators that serve as profiles that are more individualized. A profile would be a measure of where an individual stands on multiple characteristics. This progression and line of work is very difficult to do and long-term. In the meantime, treatment outcome effects can be materially improved by identifying one or two moderators.

Chapter 7

1. In case-control studies, one usually considers the cases as those individuals showing the problem or characteristic of interest and the controls as not showing that characteristic of interest. The term "healthy controls" is usually used, but there are occasional lapses in which "normal" controls is used. "Normal" is not too meaningful or helpful in this context. Apart from the methodological issue, there is the politically incorrect and insensitive issue. Use of the term "normal" to describe the control group implies that the group of cases (with the characteristic of interest) is not normal. It is likely that the case group is normal (or within the bounds of normative behaviors and characteristics), whatever that is, in all sorts of ways and hence ought not to be characterized by the feature that led to their selection in a particular study.
2. Overmentalizing has emerged in research on the theory of mind (ToM). ToM refers to the capacity to attribute mental states (e.g., thoughts, feelings, intentions, beliefs) to oneself and to others. Our mentalization—how we make sense of the world—can occur in different ways. Overmentalizing is an excessive or exaggerated style as for example that might be seen in paranoia with attributions that have gone awry. Mentalization has been studied extensively in schizophrenia research, but extended to other disorders. The broad assumption as that many of not most psychiatric disorders will involve difficulties in mentalization. Mentalization-based treatment is an intervention that specifically focuses on developing more adaptive mentalization (see Bateman & Fonagy, 2010).

Chapter 8

1. In psychological research, the designs have been referred to by different terms, such as intrasubject-replication designs, N = 1 research, and intensive designs, to mention a few. Each of the terms to describe the designs is partially misleading. For example, the terms "single-case" and "N = 1 designs" imply that only one subject is included. Often this is true, but more often multiple subjects are included. Moreover, "single-case" research occasionally includes very large groups of subjects; entire communities and cities have been included in some single-case designs (Kazdin, 2011). The term "intrasubject" is a useful term because it implies that the methodology focuses on performance of the same person over time. Yet this term too is partially misleading because some of the designs depend on looking at the effects of interventions across (i.e., between) subjects. The term intensive design has not grown out of the tradition of single-case research and is used infrequently. Also, the term "intensive" has the unfortunate connotation that the investigator is working intensively to study the subject, which probably is true but is beside the point. For purposes of conformity with many existing works, "single-case designs" is used in this chapter because it draws attention to the unique feature of the designs, i.e., the capacity to experiment with individual subjects, because it enjoys the widest use, and therefore it is the term one is most likely to encounter in reading research that uses one of the designs. (Of course, by referring to the single case, there is no intention of slighting married or cohabiting cases.)
2. The slope or gradient of the trend line can be positive (is accelerating or the line is getting higher and higher as data are collected over time—e.g., a graph that shows crime rate in a city is increasing over time) or negative (is decelerating or the line is getting lower and lower—as in a graph that shows that crime rate is decreasing). The gradient or angle of the slope reflects how fast the change is made over time, i.e., how steep the line is. There may be no increase or decrease and the line is just flat over time. Although the direction and degree of slope can be easily quantified, there is no need for that level of specificity for the present discussion.
3. Prominent in psychology single-case research designs developed out of areas of research referred to behavior analysis and includes both experimental and applied research. Key journals that publish research in this tradition and routinely use single-case designs are the *Journal of the Experimental Analysis of Behavior* and the *Journal of Applied Behavior Analysis.* Yet, the designs also appear in many other journals and across many topics (e.g., behavior and cognitive therapy, rehabilitation, and special education). In short, the designs are not restricted at all to any one area of study or

discipline, even though they are used much less frequently than between-group designs.

4. As the reader may well know, the expression "Beauty is in the eye of the beholder" is not quite accurate. Actually, research shows that there is considerable agreement in what beauty is, and who is beautiful, although there are individual taste preferences as well (e.g., Honekopp, 2006).
5. Serial dependence refers to the relation of the data points to each other in the series of continuous observations. The dependence reflects the fact that the residuals (error) in the data points are correlated (or can be) from one occasion to the next. The dependence is measured by evaluating whether the data points are correlated with each other over time (referred to as autocorrelation; see Kazdin, 2011). Serial dependence is important to know for two reasons. First, the presence of serial dependence precludes the straightforward application of statistical techniques with which we are most familiar (conventional *t* and *F* tests). Serial dependence violates a core assumption of these tests, and use of these tests gives biased estimates of the effect leading to more Type I (i.e., showing a statistically significant effect when there would not have been one) or Type II (i.e., showing no significant effect when there actually was one) errors. Second, if serial dependence exists in the data, the analysis needs to take that into account. The dependence reflects some trend or pattern in the underlying data. It may not be a simple linear trend, but a trend perhaps jolted by random effects and only detected across different lags. A data-analytic technique is needed to account for the dependence and to discern whether any intervention effect is evident over and above some overarching but possibly subtle pattern. As I noted previously, vision and visual inspection are not up to the task. I mention a solution later in the chapter.
6. Effect size and its computation were covered in Chapter 3. In terms of the magnitude of effect size, an arbitrary but widely accepted standard is to consider .2, .5, and .8 as small, medium, and large effect sizes, respectively (Cohen, 1988). As a point of reference, effect size of psychotherapy from meta-analyses of that research hovers around .7. In the context of the present discussion, requiring an effect size of 2.0 is huge and not very common.

Chapter 9

1. Positivist tradition refers to positivism as a philosophy of science that focuses on empirical evidence derived from observed experience to derive lawful relations. This tradition dominates research in the natural, biological, and social sciences. The goal is to focus on the measurable ways that are as objective and value free as possible. Underlying positivism is the notion of realism; that is, there is a real world, and the task is to describe and explain the phenomena free from the perspective, experience, or views of the researcher. Other sources of knowledge including subjective experience, introspection, and intuition are excluded from the approach as a general rule. In sharp contrast, the constructionist or interpretive tradition underscores the importance of the participants (both "subject" and "researcher") and how they perceive, consider, and experience reality. That is, reality is also a construction that is not free from the observer. This latter approach captures qualitative research where subjective experience and how individuals construct reality are central. These views have overlap, and none is the extreme my simple rendition might suggest. For example, there is a reality (e.g., exoplanets in the cosmos and one more dazzling methodology chapter comes right after this one). That is a reality and both approaches would acknowledge that. Subjective experience makes an enormous difference both in how we view the world and also in the impact of the world on us (physical and mental health). Both approaches would also agree to that, all to be elaborated in this chapter.
2. Several methods of evaluating the obtained information are available and included grounded theory methods, thematic analysis, interpretative phenomenological analysis, narrative analysis, discourse analysis, and others (see Cooper, 2012; Denzin & Lincoln, 2011). These details are beyond the scope of the chapter, which is to introduce qualitative research and its novel contributions.
3. Software for qualitative research has many options for bringing together and analyzing the data. Here are two samples (and not endorsements) (e.g., Roebuck, 2012; www.maxqda.com). For a more comprehensive and updated set of options, search "computer software for qualitative research" or equivalent terms on a Web search engine.
4. Several scientific journals are devoted to qualitative research. Examples include:
   - *Qualitative Health Research*
   - *Qualitative Inquiry*
   - *Qualitative Research in Psychology*
   - *Qualitative Social Work*
   - *Qualitative Sociology*
   - *Qualitative Studies in Education*
   - *Culture, Medicine, and Psychiatry*

   Also, an extensive list has been prepared to include journals that do not focus specifically on qualitative research but do consider and accept such research (www.slu.edu/organizations/qrc/QRjournals.html). Focusing specifically on psychology, only one of many disciplines involved in qualitative research, may convey the scale of the emphasis on quantitative rather than qualitative research. The two major psychological associations located in the United States but with international membership (e.g., American Psychological Association, Association for Psychological Science) publish over 80 journals (at the time of this writing). Only one journal (entitled, *Qualitative Psychology*) is devoted to qualitative research and began in 2014 with its first issue. In principle all of the other journals might include a qualitative study here and there, but in practice such research is not common.
5. In 2007, two journals (Journal of Mixed Methods Research and International Journal of Multiple Research Approaches) began and provided an outlet for this type of research. In the latter journal, a special issue, entitled, "Mixed Methods Research: Philosophy, Policy and Practice in Education" was published (2013, Volume 7) and provides a useful sample of research. Apart from journal publications, there is an annual international conference on mixed methods (www.methodspace.com/group/mixedmethodsresearchers/forum/topics/start-an-international) and scores of YouTube videos to describe the basics. I mention this to convey that the mixed-methods research has considerable professional interest.

Chapter 10

1. Psychological testing is a topic that goes beyond our focus on measures as tools for research. Measures are used for screening, diagnosis, selection, and placement of individuals and in many different contexts and settings (e.g., schools, clinics, business and industry, military, athletics). The Standards for Educational and Psychological Testing has been developed to address a variety of issues (e.g., selection of methods, ethical issues) (www.apa.org/science/programs/testing/standards.aspx). The most recent version of the standards was developed in 1999 jointly by three organizations (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education). The standards are not intended to address the range of methodological issues and considerations raised in research methodology. Yet, the standards are essential for those involved in testing and the application of test results well beyond the context and topic of this chapter. (For an excellent summary of the standards, see http://spb.ca.gov/content/laws/selection_manual_appendixf.pdf.) Ethical issues and treatment of participants include and go beyond assessment and are taken up later in the text.

2. With Pearson product-moment correlation (or r), it is important not to confuse statistical significance and magnitude of the effect or correlation. A statistically significant correlation (e.g., r = .20) may not be very high or reflect a very strong effect and a fairly high correlation (e.g., r = .70) may not be statistically significant. Significance of r depends heavily on sample size. There are different ways to evaluate the magnitude of the effect of r, i.e., what it means. A frequently used index is to square the correlation so that an r of .2 equals an r 2 of .04. This latter number can be made into a percentage (by multiplying by 100), and we can say that the r 2 means that 4% of the variance is shared (overlaps with is common) between the two measures that are correlated. Obviously 4% is not very strong a relation. One cannot judge the importance of a relation only by r or shared variance (overlap). For theoretical or applied reasons, even a very small relation might be critical (e.g., as a proof of concept we discussed earlier or in adding an increment in predictability in identifying suicidal adolescents).
3. There are many ways to analyze the results of multitrait-multimethod matrices to identify the extent to which trait and method variances contribute to the results (e.g., structural equation modeling, confirmatory factor analysis, multilevel modeling, generalizability theory, and others). These are beyond the scope of the chapter but are discussed and nicely illustrated elsewhere (e.g., Hox & Balluerka, 2009; Woehr, Putka, & Bowler, 2012).

Chapter 11

1. At the time of national elections, surveys tell us how a segment of the population (e.g., from one political party or another or in a given region or state in the United States) views a candidate or a critical issue. The survey data are presented to the public as if they represent "true" results. They are "true" assuming they were accurately scored and analyzed. What is not conveyed is that a differently worded survey changing how the questions were asked, the response format, and the ordering of the items might well change the "truth." Surveys play critical roles in psychological and national research. The methodological point is that scores on a measure are in part a function of characteristics of the measure, which includes wording of the items but also modality of assessment (e.g., self-report, others' report). In a study that uses all self-report measures, correlations between the measures may be due in part to the fact that they share a common method, namely, they are all based on self-report.
2. At the MTurk (Amazon Mechanical Turk; www.mturk.com/mturk/) and Qualtrics (www.qualtrics.com/) are now fairly commonly used platforms to conduct studies and run subjects via the Internet. Please see the respective Web sites for further details.

Chapter 12

1. The term "clinical significance" is used extensively in clinical psychology. A preferred term might well be applied significance or something that is broader. The reason of course is that we care about impact of our interventions (e.g., in education, safe sex, nutrition) in many contexts that are not "clinical," i.e., are not related to patient samples and mental or physical health per se.
2. *Moderator* refers to some characteristic that influences the direction or magnitude of the relation between the intervention and outcome. If the effectiveness of an intervention varies as a function of ethnicity or sex, these variables are moderators. *Mediator* is a construct that shows a statistical relation between an intervention and outcome. This is an intervening construct that suggests processes about why change occurs or on which change depends.

Chapter 13

1. Many statistical tests (e.g., factor analysis, regression, cluster analyses, time-series analysis, path analyses) include a number of decision points about various solutions, parameter estimates, cutoffs for including or deleting variables in the analysis or model, and so on. These decisions often are not made by the investigator but are accepted by the "default" criteria in the data-analytic programs. Independently of who makes the decision, there are assumptions and sources of subjectivity in the decision that can greatly influence the yield from statistical tests.
2. An interesting and very readable discussion of how these p levels came to be adopted and hence why they ought to be viewed quite tentatively is available elsewhere (Cowles & Davis, 1982). That article conveys that conventional levels of .05 and .01 are rather arbitrary. Early in my career—actually when analyzing the results of my dissertation—I began to abandon $p < .05$ and adopted $p < .33$ as *the* level for significance. Through a rather amazing coincidence, most of the hypotheses of my dissertation were supported at $p < .33$. The bulk of my dissertation orals was spent by my committee quibbling with my attorney about my right to adopt this level for alpha (e.g., the U.S. Constitution is rather hazy on individual, state, and federal rights in selecting alpha) and whether I could be charged with impersonating a scientist.
3. Effect size (ES) here will be used to refer to the mean difference between two groups divided by the standard deviation. This is also Cohen's d.
4. There is more than one type of Bayesian analyses (e.g., Bayes factor approach, parameter estimation, hierarchical Bayesian analysis) all beyond the present scope. There are excellent introductory resources for explaining Bayesian analyses and its underpinnings but also guides for use of software and computation (see Kruschke, 2011b). In addition, there is a journal called *Bayesian Analysis* published by the International Society for Bayesian Analysis. It seeks to publish a wide range of articles that demonstrate or discuss Bayesian methods in some theoretical or applied context.

Chapter 14

1. The techniques of EDA are summarized by the use of four techniques or the "4 Rs" (Behrens & Yu, 2003). These include *Revelation* through the use of graphics, *Re-expression* of the data through scale transformation, *Residuals* by using model building and measures to understand their structure, and *Resistance* that refers to being insensitive to many perturbations in the data (e.g., by using ranking of subjects and medians, which are less sensitive to some sources of variability in the data). As one can see, EDA is a formal approach to data exploration. The procedures are technical and, lamentably, not usually included in undergraduate and graduate training in methodology. Hence, they are omitted from the present text. Further reading provides options for the interested reader, including several software programs that are much less familiar than the more commonly used packages (e.g., SPSS) (see Behrens, DiCerbo, Yei, & Levy, 2013).

Chapter 15

1. The Kazdin–Nock Illusion is a variant of the more familiar Figure/Ground Illusion (as depicted in the vase/profile figure most readers will know) (Kazdin & Nock, 2003). In relation to statistical analyses and interpretation, the K–N Illusion works like this. An investigator sees arrows plotted from a data analysis or chart. These arrows point in a particular direction between one or more "predictors" and an outcome. The data analyses and the arrows suggest a direction whether or not a time line actually permits one to infer that the predictor came before the "outcome." The investigator *figures* that these arrows are good *grounds* for concluding a causal relation, ergo the resemblance to the Figure/Ground Illusion. This is an illusion.
2. There are a number of solutions to the concern and impact of the publication bias. One of them is to provide a forum for publishing negative results. The *Journal of Articles in Support of the Null*

Hypothesis (www.jasnh.com/) provides a free online journal that covers all areas of psychology. The journal offers "an outlet for experiments that do not reach the traditional significance levels ($p < .05$)." (p. 1, Web site). The opening statement further notes, "Without such a resource researchers could be wasting their time examining empirical questions that have already been examined." Outside of psychology other journals share the same goal. Two examples are the *Journal of Negative Results in BioMedicine* (www.jnrbm.com) and *The All Results Journals* (www.arjournals.com/ojs/), which cover a few different disciplines (e.g., nanotechnology, biology, physics). Each of these journals publishes "negative results" and helps redress the publication bias. Yet, the solution to publish more negative results has not caught on heavily within social, biological, or natural science.

3. The Institute of Medicine (IOM) is an independent, nonprofit and nongovernment organization that http://resources.iom.edu/widgets/timeline/index.html? keepThis=true&TB_iframe=true&height=710&width=1000; is designed to provide unbiased and authoritative advice to decision makers and the public. In 1863, President Abraham Lincoln established the National Academy of Sciences to provide any department of government with expertise on any subject. Experts on the topic usually from diverse disciplines are convened to evaluate a given area. Members receive no financial report. The IOM is part of the National Academies and focuses on issues of health and health care. The IOM as other branches of the National Academies provides information in response to mandates from Congress, other federal agencies, or independent organizations. Many reports are issued on a range of topics (e.g., health care, nutrition, climate change, and health) and provide informed evaluations of what is known on the topic at a given point in time (see www.iom.edu/About-IOM.aspx).
4. As you recall, Ivan Pavlov (1849–1936) elaborated respondent or classical conditioning, which refers to how associations are made between stimuli and reflex responding. You have read that a sound, light, or signal can be made to elicit a reflex response (e.g., salivation, startle) by the special pairing of these unlearned stimuli with the actual stimuli that elicit the behavior. Skinner elaborated operant conditioning, which focuses on behaviors that operate in the world (walking, talking, doing homework) and how these behaviors are influenced and can be developed.
5. I have emphasized the R Project that focuses on replication in psychology. As noted here, the concerns and renewed priority of replication research spans many areas. For example, there is a Reproducibility Initiative in medicine and in more focused areas within that (e.g., cancer research) (see Couzin-Frankel, 2013a; Laine, Goodman, Griswold, & Sox, 2007).

Chapter 16

1. The use of the Internet (e.g., e-mail, social media) requires the transfer of information from one computer to another. To do this, each computer requires an Internet Protocol Address (or IP address), which is personally identifiable information that is automatically registered when any communication is made over the Internet (e.g., visiting any Web site, sending or receiving messages). The IP address can be connected with one's browsing history and routinely is sent to third parties (e.g., other Websites that track behavior). Collection of IP addresses alone might not be considered as an invasion of privacy, but the address can be associated with all sorts of activities to which individuals are unaware and for which they have not provided consent. Increasingly research uses the Internet as a means of collecting data (e.g., Amazon Mechanical Turk [MTurk], Qualtrics). Additional protections are needed in cases where the information might be viewed as private. In some cases (e.g., with patients), special encryption of messages over the Internet is required.
2. MTurk provides a Web services system that allows one to obtain and run subjects who receive money for their efforts. Many studies are now run in which people with access to the Internet can elect to participate in experiments (see www.mturk.com). Qualtrics is private software company that also provides the opportunity to collect data from the Web as well as providing other services such as statistical analyses (see http://qualtrics.com/research-suite/#enterprise). Increasingly Web-based assessments are being conducted because obtaining large numbers of subjects (e.g., hundreds) can be rapid (few days) and the process is more streamlined and efficient than recruiting introductory psychology students.
3. Polio is a disease that mainly affects children under 5 years of age. The infection can lead to irreversible paralysis (usually in the legs) and for 5–10% who suffer death (when their muscles to breathe become immobilized). A live oral polio vaccine is used worldwide especially in countries where polio is more common. The oral vaccine, in comparison to the injected inactivated poliovirus vaccine, is used because it is less expensive and easier to administer, can protect entire communities that are critical for eradication, and does not require trained people (e.g., nurses) to administer injections. In most circumstances, the vaccine produces a harmless infection in the intestines and builds resistance and protects against polio. Yet, a rare side effect is contracting polio. In the United States, injections are given and in that version, there is no active virus and does cause polio and paralysis as side effects. The goal for complete eradication of polio includes elimination of the oral vaccination (Aylward & Yamada, 2011; Orenstein, 2013).
4. The Tuskegee Syphilis Study often is routinely presented to convey critical ethical issues in research. It is important to know the study for several reasons including ethical breaches, racism and discrimination, but also broader issues such as the critical role of oversight for all that we do. As researchers, one occasionally claims or feels, "why am I going through all these hoops for the Institutional Review Board, subject protections, and so on." The Tuskegee study, experiments of Nazis during the war, but other studies as well convey that flagrant violations of humane codes can lead to cruel treatment and death. For example, an extremely influential paper published in 1966 (and republished in 1976) reported on ethical violations among over 20 researchers and their publication in major (prestigious) journals (e.g., *New England Journal of Medicine, Science*) (Beecher, 1966). The violations include flagrant examples (e.g., withholding antibiotics from men with rheumatic fever, injecting live cancer cells into nursing home patients). This paper and other similar work at about the same time (Pappworth, 1967) were important to convey that ethical lapses and mistreatment of subjects (e.g., no informed consent, not conveying risks) of the Tuskegee study are not restricted to horrendous lapses during war by demonic regimes. Rather they were more common and reflected in situations where the highest standards of research supposedly were invoked. Eventually through other media, word of the Tuskegee study reached the public and concerns were voiced about "human guinea pigs." The work greatly influenced the development of ethical guidelines and oversight of scientific research (Edelson, 2004; Harkness, Lederer, & Wikler, 2001).
5. The WMA, founded in 1947, is an organization that represents the interests of over 9 million physicians and in over 100 countries. The goal was to develop an organization to ensure the independence of physicians and their work in meeting the highest possible standards for ethical behavior at all times. The deep concern at that time emerged at the end of World War II and the use of physicians as part of medical atrocities and experiments of the Nazi regime. The broad goal is to establish and promote standard of ethical behavior and care by physicians. Committees within the WMA are convened to make policy statements are made on a variety of issues, beyond the research foci emphasized in the present chapter, including public health (e.g., the importance of vaccination against influence), human rights (e.g., condemnation of torture), and many other such issues (e.g., children's right to health). Publications, policy statements, educational resources, and media contacts are used to convey the policies. None of the

policies or positions are legally binding (e.g., whether and when to use placebo control conditions), but the statements can actively influence policies of countries and research institutions that do have binding rules and guidelines (see www.wma.net/en/10home/index.html).

Chapter 17

1. The Cochrane Collaboration is an international organization that spans more than 120 countries. The goal of the organization is to provide high-quality and unbiased reviews on topics related to health care and health policy. These include interventions for prevention, treatment, and rehabilitation. Many of the Cochrane Reviews, as the products are called, are made available and published online. As noted on the Web page (www.cochrane.org/about-us), the organization is named after Archie Cochrane (1909–1988), a British epidemiologist, who advocated the use of randomized controlled trials as a means of reliably informing healthcare practice.
2. Stapel has written a textbook, only in Dutch at the time of this writing in which he provides a detailed account of the fraud and its consequences (Stapel, 2012). A review of this textbook is available in English (Borsboom & Wagenmakers, 2013).
3. SafeAssign is one of many software services to help check and guide students to avoid plagiarism (see www.safeassign.com/). The checking of a paper or proposal is made by comparing what was one of us has written to large databases, including documents publically available on the Internet, over 1,100 publication titles and approximately 2.6 million articles in other databases, and others. Copied or suspicious text passages are so identified.
4. Increasingly research is collaborative and with that novel issues emerge in allocation of credit. For example, it is not rare for an article to have 100 or more authors (e.g., elaborating the genome). The most extreme case I could identify was an article that included more than 37,000 authors as a product of online research (using crowdsourcing) in biochemistry (Lee et al., 2014). More likely for clinical psychology a small set of authors (e.g., 3–6) will prepare an article, although more authors might be included occasionally (e.g., 10+ authors). In principle, these articles do not raise scientific integrity issues of a special nature beyond those discussed in this section. Large-scale collaborations do raise other issues professionally such as challenges in allocating credit and considering the studies in relation to the promotion of individuals based on their publication record. These issues are beyond the scope of this chapter.
5. The topic of big data is enormous and clearly a wave that will affect all of the sciences. Indeed, phrases such as the "Era of big data" (Bollier, 2010, p. 1) and "big data revolution" (Kolb & Kolb, 2013, title) are used to describe the movement. A new field of study is discussed (data science) as a multidisciplinary discipline to bring to bear the diverse sources of expertise in programming, math, software, technology, and more (Nielsen & Burlingame, 2012). New research centers devoted to big data have been formed and encouraged. For example, the National Institutes of Health (2013d) has provided funds to foster the development of such centers.
6. Conflict of interest emerges in other contexts than research, such as multiple role relations with current and former clients in the context of psychotherapy (APA, 2010a). These are important but not the main issues that arise in research and beyond the scope of this chapter; the reader is referred elsewhere (Welfel, 2013).

Chapter 18

1. Reporting of ethnic composition of the sample is a standard practice in psychological studies within the United States. In some other countries (e.g., France, Canada), asking, seeking this information from clients and reporting it are considered discriminatory and the information is not available to report.
2. Preparing a manuscript for publication entails several format requirements, such as print style and size, citations of sources, use of abbreviations, structure of tables and figures, and order in which sections of the article appear. These are detailed in the *Publication Manual of the American Psychological Association* (APA, 2010b) and are not covered in this chapter. Also, studies are reported in other formats than manuscript or journal form. For example, poster sessions are one format often used for presenting one's work, especially early in one's career. Here too there are many concrete and practical issues in constructing posters and excellent sources (see For Further Reading). This chapter focuses on methodological thinking and underpinnings of communicating one's science and therefore eschews many of the practical tasks nicely covered elsewhere.
3. At the time of this writing, I am editor of *Clinical Psychological Science,* so mention of this journal could easily be construed promoting one journal over another and a conflict of interest. Mentioning or promoting the journal does not lead to any financial gain on my part.
4. A quantitative measure to evaluate journals is referred to as the "impact factor," and is based on the frequency with which articles in the journal in a given time period (2 years) in proportion to the total number articles published in the journal. An objective quantitative measure of impact has multiple uses for different parties who have interest in the impact of a journal (e.g., libraries making subscription decisions, publisher evaluating the status of a particular journal it has published). Administrators and faculty peers often use impact of the journals in which a colleague publishes as well as how often their work is cited by others among the criteria used for job appointments and promotions in academic rank, and salary adjustments. There has been a strong movement to no longer use impact factor as a way to evaluate research or merit of an investigator conducting that research (see Alberts, 2013). Impact was not designed to measure that and is subject to all sorts of influences (e.g., that vary by discipline, artifacts of publishing practices of individual journals) and that impact factor bears little relation to expert views of scientific quality. In 2012, an organization (San Francisco Declaration of Research Assessment, abbreviated as DORA), initiated at a meeting of the American Society for Cell Biology and including many editors and publishers examined the ways in which journals are evaluated. Among the consequences was agreement that "impact factor" might be useful for the purposes for which it was intended, but not for evaluating the merit of scientific research. Consequently DORA was urging journals and scientific organizations to drop the use of impact factor as an index of quality of the journal or articles in which the journal appears. Now many scientific and professional organizations (>400 at the time of this writing) and researchers (~1,000) have signed on to this recommendation to not use or flaunt impact factor as an index of quality (http://am.ascb.org/dora/). Even so, many journals still flaunt their "impact factor." It is important to mention here in case the reader is considering this as a main or major reason for submitting a manuscript to one journal rather than another.
5. Excellent readings are available to prepare the author for the journal review process (*The Trial* by Kafka, *The Myth of Sisyphus* by Camus, and *Inferno* by Dante). Some experiences (e.g., root canal without an anesthetic, income tax audit, identity theft) also are touted to be helpful because they evoke reactions that mimic those experienced when reading reviews of one manuscript. Within clinical psychology, various conceptual views (e.g., learned helplessness), clinical disorders (e.g. posttraumatic stress disorder), and intervention strategies (e.g., stress management, anger control training) are helpful in understanding and preparing oneself for negotiating its shoals.
6. The suspense as to whether one's manuscript will be accepted or rejected for publication has been addressed in a novel journal referred to as the *Journal of Universal Rejection* (www.universalrejection.org/). As the opening Web page notes, "You can send your manuscript here without suffering waves of anxiety regarding the eventual fate of your

submission. You know with 100% certainty that it will not be accepted for publication." The Web site lists prior years of journal issues and their table of contents; each issue is empty because manuscripts are never accepted. A novel idea to be sure and although not serious may provide good training for new authors as they submit their works.

7. Thanks to my dissertation committee again for letting me quote from their comments.

Chapter 19

1. This chapter provides a perspective and broad comments on methodology and where it fits in science. It is important to highlight these broad issues in part to convey that substantive advances rely heavily on the methods we have discussed in this text. At the same time, I understand the need to be of concrete help in designing a study. For persons beginning a research project, the broad issues are of little help. The initial question is where to begin? The text has moved from such topics as the sources of ideas, how to translate them to hypotheses and operational definitions, and so on. Yet, in a world of fast food and instant communication and posting material on our social media, is there something I can provide that will help the interest reader, investigator, and new scientist quickly design a methodologically wonderful study? The end of this chapter includes an Appendix to provide such a tool.
2. There is increased recognition of the importance of methodological diversity as evident by journals that foster the combination different research methods and traditions such as quantitative and qualitative research (e.g., Journal of Mixed Methods, International Journal of Mixed Methods in Applied Business and Policy Research). Other journals are even more explicit about their openness to diversity of research approaches (e.g., Multiple Research Approaches). For a given field or discipline (e.g., clinical psychology, education) and for sciences (social, biological, natural) journals that promote methodological diversity are not mainstream publication outlets that are among the most widely recognized. In addition, in training of graduate students, few programs teach multiple research traditions and methods. Yet, the message of the text is that in studying a phenomenon and pursuing an area of interest, try to draw on methods (e.g., assessments, designs, evaluation techniques) that go beyond the usual methods used in the areas in which one is working. Collaboration with others is one means of expanding horizons in ways that can greatly extend what one learns from a study.

Credits

Chapter 2 Page 28: The illusion is named after the person who is credited with its identification 1889 by a German sociologist named Franz Carl Müller-Lyer (1857–1916).

Chapter 3 Page 70: Nelson, J. C., & Devanand, D. P. (2011, p. 577). A systematic review and meta-analysis of placebo-controlled antidepressant studies in people with depression and dementia. Journal of the American Geriatrics Society, 59(4), 577–585; Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011, p. 1105). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience, 14(9), 1105–1107; Page 72: Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011, p. 1105). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience, 14(9), 1105–1107.

Chapter 4 Page 88: Bisakha, S. (2010, p. 187). The relationship between frequency of family dinner and adolescent problem behaviors after adjusting for other family characteristics. Journal of Adolescence, 33(1), 187–196.; page 91: Paul, G.L. (1967, p. 111). Outcome research in psychotherapy. Journal of Consulting Psychology, 31, 109–118.; page 93: http://en.wikipedia.org/wiki/Mediation_(statistics); page 97: http://www.cc.nih.gov/ccc/btb/ ; page 106: Garcia, J. R., Reiber, C., Massey, S. G., & Merriwether, A. M. (2012, p. 161). Sexual hookup culture: A review. Review of General Psychology, 16(2), 161–176.

Chapter 5 Page 128: Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research and teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally.

Chapter 6 Page 150: Church, R.M. (1964). Systematic effect of random error in the yoked control design. Psychological Bulletin, 62, 122–131.

Chapter 8 Page 199: Ahearn, W.H., Clark, K.M., MacDonald, R.P.F., & Chung, B.I. (2007). Assessing and treating vocal stereotypy in children with autism. Journal of Applied Behavior Analysis, 40, 263–275; page 203: Musser, E.H., Bray, M.A., Kehle, T.J., & Jenson, W.R. (2001). Reducing disruptive behaviors in students with serious emotional disturbance. School Psychology Review, 30, 294–304; page 204: Cunningham, T.R., & Austin, J. (2007). Using goal setting, task clarification, and feedback to increase the use of the hands-free technique by hospital operating room staff. Journal of Applied Behavior Analysis, 40, 673–677; page 208: Allen, K. D., & Evans, J. H. (2001). Exposure-based treatment to control excessive blood glucose monitoring. Journal of Applied Behavior Analysis, 34, 497–500; page 209: Flood, W.A., & Wilder, D.A. (2004). The use of differential reinforcement and fading to increase time away from a caregiver in a child with separation anxiety disorder. Education and Treatment of Children, 27, 1–8; page 217: Quesnel, C., Savard, J., Simard, S., Ivers, H., & Morin, C.M. (2003). Efficacy of cognitive-behavioral therapy for insomnia in women treated for nonmetastatic breast cancer. Journal of Consulting and Clinical Psychology, 71, 189–200.

Chapter 9 Page 221: Denzin, N.H. & Lincoln, Y.S. (Eds.). (2011, p. 2). The Sage handbook of qualitative research (4th. ed). Thousand Oaks, CA: Sage.

Chapter 10 Page 253: Satcher, D. (2001, v). Department of Health and Human Services (2001). Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the Surgeon General. Washington, DC: U.S. Department of Health and Human Services; Lewis-Fernández, R., & Díaz, N. (2002). The cultural formulation: A method for assessing cultural factors affecting the clinical encounter. Psychiatric Quarterly, 73(4), 271–295; page 268: Kazdin, A.E., French, N.H., Unis, A.S., Esveldt-Dawson, K., & Sherick, R.B. (1983). Hopelessness, depression and suicidal intent among psychiatrically disturbed inpatient children. Journal of Consulting and Clinical Psychology, 51, 504–510.

Chapter 11 Page 287: Data from J.S. Comer & P.C. Kendall (Eds.). The Oxford handbook of research strategies for clinical psychology (pp. 188–209). New York: Oxford University Press. page 176; page 293: Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (2000). Unobtrusive measures (revised edition). Thousand Oaks, CA: Sage Publications.

Chapter 12 Pages 316: Wolf, M.M. (1978). Social validity: The case of subjective measurement or how applied behavior analysis is finding its heart. Journal of Applied Behavior Analysis, 11, 203–214; page 317: Substance Abuse and Mental Health Services Administration (2011). SAMHSA announces a working definition of "recovery" from mental disorders and substance abuse disorders. Available on line at www.samhsa.gov/newsroom/advisories/1112223420.aspx.

Chapter 13 Page 326: Ancient Greeks, Aristotle; Tukey, J.W. (1991, p. 100). The philosophy of multiple comparisons. Statistical Science, 6, 100–116; Cohen, J. (1990, p. 1308). Things I have learned (so far). American psychologist, 45(12), 1304–1312; page 327: Mosteller, F. (2010, p. 227). The pleasure of statistics: The autobiography of Frederick Mosteller. S.E. Feinberg, D.C. Hoaglin, & J.M. Tanur (Eds.). New York: Springer; page 331: Data from Cohen, J. (1988). Statistical power analysis in the behavioral sciences. (2nd ed.). pp. 36–37. Hillsdale, NJ: Erlbaum. Reprinted with permission; page 335: Ruxton, G.D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? Methods in Ecology and Evolution, 1, 114–117; page 337: Kirk, R.E. (1996, p. 747). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746–759; Meehl, P. (1978, p. 817). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806–834; page 338: Rosnow, R.L., & Rosenthal, R. (1989, p. 1277). Definition and interpretation of interaction effects. Psychological Bulletin, 105, 143–146; Tukey, J.W. (1991,

p. 100). The philosophy of multiple comparisons. Statistical Science, 6, 100–116.

Chapter 14 Page 336: These steps were obtained from Donnellan, M.B., & Lucas, R.E. (2013). Secondary data analysis. In T.D. Little (Ed.), The Oxford handbook of quantitative methods (Vol. 2, pp. 665–677). New York: Oxford University Press.

Chapter 15 Page 373: Fisher, L. B., Miles, I. W., Austin, S. B., Camargo Jr, C. A., & Colditz, G. A. (2007, p. 7, on line). Predictors of initiation of alcohol use among US adolescents: Findings from a prospective cohort study. Archives of Pediatrics & Adolescent Medicine, 161(10), 959–966; page 385: Snapinn, S., Chen, M. G., Jiang, Q., & Koutsoukos, T. (2006). Assessment of futility in clinical trials. Pharmaceutical Statistics, 5(4), 273–281; page 389: Schmidt, S. (2009) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology,13(2), 90–100; page 390: Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012, p. 633 for more on this quote). An agenda for purely confirmatory research. Perspectives on Psychological Science, 7(6), 632–638; page 393: The Journal of Articles in Support of the Null Hypothesis (www.jasnh.com/). p. 1, web site.

Chapter 16 Page 408: The Sample Paragraph was adapted from University of Virginia (2013, ww.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html) University of Virginia (2013); Institutional Review Board for Social and Behavioral Sciences. Sample briefing statement. Retrieved from www.virginia.edu/vpr/irb/sbs/resources_guide_deception_debrief_sample.html) Copyright 2013 by the Rector and Visitors of the University of Virginia; page 416: US Department of Health and Human Services (2009; www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.116) ; page 418: http://grants.nih.gov/grants/policy/coc/ , quote from the web site, National Institutes of Health (2013a). Certificate of confidentiality. Available on line at http://grants.nih.gov/grants/policy/coc/ ; http://grants.nih.gov/grants/policy/coc/faqs.htm#365 ; page 426: **Table 16–6 & 16–07**:Copyright © 2010 by the American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is [list the original APA bibliographic citation]. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

Chapter 17 Page 432: Albert Einstein quoted in National Research Council. (2002, p. 16). Integrity in scientific research: Creating an environment that promotes responsible conduct. Washington, DC: National Academy Press. http://books.nap.edu/catalog.php?record_id=10430#toc ; page 433: Copyright © 2010 by the American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is [list the original APA bibliographic citation]. No further reproduction or distribution is permitted without written permission from the American Psychological Association; page 437: Several of the points in this table have been discussed by others (e.g., John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23(5), 524–532.; Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366.); page 442: National Institutes of Health (2013e). NIH policy on mitigating risks of life sciences dual use research of concern. Notice Number: OT-OD-13-107. Bethesda, MD: NIH. Available on line at. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-107.html ; page 440: Vieyra, M., Strickland, D., & Timmerman, B. (2013, p. 39). Patterns in plagiarism and patchwriting in science and engineering graduate students' research proposals. International Journal for Educational Integrity, 9(1), 35–49; page 443: Eggert, L. D. (2011, Table 1). Best practices for allocating appropriate credit and responsibility to authors of multi-authored articles. Frontiers in psychology, 2. Available on line at www.ncbi.nlm.nih.gov/pmc/articles/PMC3164109/. Copyright 2011 Egger; International Committee of Medical Journal Editors. (2013). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: Roles and responsibilities of authors, contributors, reviewers, editors, publishers, and owners: Defining the role of authors and contributors. www.icmje.org/roles_a.html ; page 450: National Research Council (2004, p. 2). Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology. Biotechnology research in an age of terrorism. Washington, DC: National Academy Press; page 454: Guyatt, G., Akl, E. A., Hirsh, J., Kearon, C., Crowther, M., Gutterman, D., . . . & Schnemann, H. (2010). The vexing problem of guidelines and conflict of interest: A potential solution. Annals of Internal Medicine, 152(11), 738–741; page 455: Bohannon, J. (2013, p. 62). Who's afraid of peer review? Science, 342, 60–65; page 460: www.cochrane.org/about-us .

Chapter 18 Page 475: The Journal of Universal Rejection (www.universalrejection.org/); page 472: www.consort-statement.org ; page 473: Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., . . . & Barlow, D. H. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A randomized controlled trial. Behavior Therapy, 43(3), 666–678.

Name Index

C

D

E

F

G

K

L

M

N

O

P

Q

R

S

T

Subject Index

N

O

P

Q

R