OXFORD LIBRARY OF PSYCHOLOGY

EDITED BY

JONATHAN S. COMER &

PHILIP C. KENDALL

The Oxford Handbook *of* RESEARCH STRATEGIES *for* CLINICAL PSYCHOLOGY

The Oxford Handbook of Research Strategies for Clinical Psychology

OXFORD LIBRARY OF PSYCHOLOGY

EDITOR-IN-CHIEF

Peter E. Nathan

AREA EDITORS:

Clinical Psychology
David H. Barlow

Cognitive Neuroscience
Kevin N. Ochsner and Stephen M. Kosslyn

Cognitive Psychology
Daniel Reisberg

Counseling Psychology
Elizabeth M. Altmaier and Jo-Ida C. Hansen

Developmental Psychology
Philip David Zelazo

Health Psychology
Howard S. Friedman

History of Psychology
David B. Baker

Methods and Measurement
Todd D. Little

Neuropsychology
Kenneth M. Adams

Organizational Psychology
Steve W. J. Kozlowski

Personality and Social Psychology
Kay Deaux and Mark Snyder

OXFORD LIBRARY OF PSYCHOLOGY

Editor in Chief PETER E. NATHAN

The Oxford Handbook of Research Strategies for Clinical Psychology

Edited by

Jonathan S. Comer

Philip C. Kendall

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press in the UK and
certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
The Oxford handbook of research strategies for clinical psychology / edited by Jonathan S. Comer, Philip C. Kendall.
pages ; cm
Includes bibliographical references.
ISBN 978–0–19–979354–9
1. Clinical psychology—Research—Methodology—Handbooks, manuals, etc. I. Comer, Jonathan S.,
editor of compilation. II. Kendall, Philip C., editor of compilation. III. Title: Handbook of
research strategies for clinical psychology.
RC467.8.O94 2013
616.890072—dc23
2012034498

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

SHORT CONTENTS

This page intentionally left blank

OXFORD LIBRARY OF PSYCHOLOGY

The *Oxford Library of Psychology,* a landmark series of handbooks, is published by Oxford University Press, one of the world's oldest and most highly respected publishers, with a tradition of publishing significant books in psychology. The ambitious goal of the *Oxford Library of Psychology* is nothing less than to span a vibrant, wide-ranging field and, in so doing, to fill a clear market need.

Encompassing a comprehensive set of handbooks, organized hierarchically, the *Library* incorporates volumes at different levels, each designed to meet a distinct need. At one level are a set of handbooks designed broadly to survey the major subfields of psychology; at another are numerous handbooks that cover important current focal research and scholarly areas of psychology in depth and detail. Planned as a reflection of the dynamism of psychology, the *Library* will grow and expand as psychology itself develops, thereby highlighting significant new research that will impact on the field. Adding to its accessibility and ease of use, the *Library* will be published in print and, later on, electronically.

The *Library* surveys psychology's principal subfields with a set of handbooks that capture the current status and future prospects of those major subdisciplines. This initial set includes handbooks of social and personality psychology, clinical psychology, counseling psychology, school psychology, educational psychology, industrial and organizational psychology, cognitive psychology, cognitive neuroscience, methods and measurements, history, neuropsychology, personality assessment, developmental psychology, and more. Each handbook undertakes to review one of psychology's major subdisciplines with breadth, comprehensiveness, and exemplary scholarship. In addition to these broadly-conceived volumes, the *Library* also includes a large number of handbooks designed to explore in depth more specialized areas of scholarship and research, such as stress, health and coping, anxiety and related disorders, cognitive development, or child and adolescent assessment. In contrast to the broad coverage of the subfield handbooks, each of these latter volumes focuses on an especially productive, more highly focused line of scholarship and research. Whether at the broadest or most specific level, however, all of the *Library* handbooks offer synthetic coverage that reviews and evaluates the relevant past and present research and anticipates research in the future. Each handbook in the *Library* includes introductory and concluding chapters written by its editor to provide a roadmap to the handbook's table of contents and to offer informed anticipations of significant future developments in that field.

An undertaking of this scope calls for handbook editors and chapter authors who are established scholars in the areas about which they write. Many of the nation's and world's most productive and best-respected psychologists have agreed to edit *Library* handbooks or write authoritative chapters in their areas of expertise.

For whom has the *Oxford Library of Psychology* been written? Because of its breadth, depth, and accessibility, the *Library* serves a diverse audience, including graduate students in psychology and their faculty mentors, scholars, researchers, and practitioners in psychology and related fields. Each will find in the *Library* the information they seek on the subfield or focal area of psychology in which they work or are interested.

Befitting its commitment to accessibility, each handbook includes a comprehensive index, as well as extensive references to help guide research. And because the *Library* was designed from its inception as an online as well as a print resource, its structure and contents will be readily and rationally searchable online. Further, once the *Library* is released online, the handbooks will be regularly and thoroughly updated.

In summary, the *Oxford Library of Psychology* will grow organically to provide a thoroughly informed perspective on the field of psychology, one that reflects both psychology's dynamism and its increasing interdisciplinarity. Once published electronically, the *Library* is also destined to become a uniquely valuable interactive tool, with extended search and browsing capabilities. As you begin to consult this handbook, we sincerely hope you will share our enthusiasm for the more than 500-year tradition of Oxford University Press for excellence, innovation, and quality, as exemplified by the *Oxford Library of Psychology.*

Peter E. Nathan
Editor-in-Chief
Oxford Library of Psychology

ABOUT THE EDITORS

Jonathan S. Comer

Dr. Comer is Associate Professor of Psychology at Florida International University and the Center for Children and Families. Before this, he served as Director of the Early Childhood Interventions Program of Boston University, an interdisciplinary clinical research laboratory in the Center for Anxiety and Related Disorders devoted to expanding the quality and accessibility of mental health care for very young children. His program of research examines five areas of overlapping inquiry: (1) the assessment, phenomenology, and course of child anxiety disorders; (2) the development and evaluation of evidence-based treatments for childhood psychopathology, with particular focus on the development of innovative methods to reduce systematic barriers to effective mental health care in the community; (3) the psychological impact of disasters and terrorism on youth; (4) national patterns and trends in the utilization of mental health services and quality of care; and (5) psychosocial treatment options for mood, anxiety, and disruptive behavior problems presenting in early childhood.

Philip C. Kendall

Dr. Kendall's doctorate in clinical psychology is from Virginia Commonwealth University. He has been honored with the Outstanding Alumnus Award from this institution. His Board Certification (ABPP) is in (a) Clinical Child and Adolescent psychology and (b) Cognitive and Behavioral Therapy. Dr. Kendall's CV lists over 450 publications. He has had over 25 years of uninterrupted research grant support from various agencies. Having received many thousands of citations per year, he placed among an elite handful of the most "Highly-Cited" individuals in all of the social and medical sciences. In a recent quantitative analysis of the publications by and citations to all members of the faculty in the 157 American Psychological Association-approved programs in clinical psychology, Dr. Kendall ranked 5th. Dr. Kendall has garnered prestigious awards: Fellow at the Center for Advanced Study in the Behavioral Sciences, inaugural Research Recognition Award from the Anxiety Disorders Association of America, "Great Teacher" award from Temple University, identified as a "top therapist" in the tristate area by *Philadelphia* Magazine, and a named chair and Distinguished University Professorship at Temple University. He has been president of the Society of Clinical Child and Adolescent Psychology (Division 53) of APA as well as President of the Association for the Advancement of Behavior Therapy (AABT, now ABCT). Recently, ABCT recognized and awarded him for his "Outstanding Contribution by an Individual for Educational/Training Activities."

This page intentionally left blank

CONTRIBUTORS

Elizabeth W. Adams
University of Alabama
Tuscaloosa, AL

Leona S. Aiken
Department of Psychology
Arizona State University
Tempe, AZ

Marc Atkins
Department of Psychiatry
University of Illinois at Chicago
Chicago, IL

Amanda N. Baraldi
Department of Psychology
Arizona State University
Tempe, AZ

David H. Barlow
Center for Anxiety and Related Disorders
Boston University
Boston, MA

Rinad S. Beidas
Center for Mental Health Policy and Services Research
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA

Deborah C. Beidel
Psychology Department
University of Central Florida
Orlando, FL

Timothy A. Brown
Department of Psychology
Boston University
Boston, MA

Mathew M. Carper
Department of Psychology
Temple University
Philadelphia, PA

Heining Cham
Department of Psychology
Arizona State University
Tempe, AZ

Candice Chow
Department of Psychology
Wellesley College
Wellesley, MA

Jonathan S. Comer
Department of Psychology
Florida International University
Miami, FL

Mark R. Dadds
School of Psychology
The University of New South Wales
Sydney, Australia

Ulrich W. Ebner-Priemer
Karlsruhe Institute of Technology
Karlsruhe, Germany

Andy P. Field
School of Psychology
University of Sussex
Sussex, United Kingdom

John P. Forsyth
University at Albany, State University of New York
Department of Psychology
Albany, NY

Kaitlin P. Gallo
Center for Anxiety and Related Disorders
Boston University
Boston, MA

Lois A. Gelfand
Department of Psychology
University of Pennsylvania
Philadelphia, PA

Andrew J. Gerber
Division of Child and Adolescent Psychiatry
Columbia College of Physicians and Surgeons
New York State Psychiatric Institute
New York, NY

Marlen Z. Gonzalez
Department of Psychology
University of Virginia
Charlottesville, VA

David J. Hawes
School of Psychology
The University of Sydney
Sydney, Australia

Nadia Islam
Department of Psychology
Virginia Commonwealth University
Richmond, VA

Matthew A. Jarrett
Department of Psychology
University of Alabama
Tuscaloosa, AL

Kirsten Johnson
The University of Vermont
Burlington, VT

Zornitsa Kalibatseva
Department of Psychology
Michigan State University
East Lansing, MI

Philip C. Kendall
Department of Psychology
Temple University
Philadelphia, PA

Gerald P. Koocher
Department of Psychology
Simmons College
Boston, MA

Helena Chmura Kraemer
Stanford University
University of Pittsburgh
Pittsburgh, PA

Frederick T. L. Leong
Department of Psychology
Michigan State University
East Lansing, MI

Yu Liu
Arizona State University
Tempe, AZ

Ginger Lockhart
Department of Psychology
Arizona State University
Tempe, AZ

David P. MacKinnon
Department of Psychology
Arizona State University
Tempe, AZ

Katherine M. McKnight
Department of Psychology and Pearson Education
George Mason University
Fairfax, VA

Patrick E. McKnight
Department of Psychology and Pearson Education
George Mason University
Fairfax, VA

Bryce D. McLeod
Department of Psychology
Virginia Commonwealth University
Richmond, VA

Tara Mehta
Department of Psychiatry
University of Illinois at Chicago
Chicago, IL

Jenna Merz
Department of Psychiatry
University of Illinois at Chicago
Chicago, IL

Dave S. Pasalich
School of Psychology
The University of New South Wales
Sydney, Australia

Joseph R. Rausch
Department of Pediatrics
University of Cincinnati
Cincinnati, OH

Kendra L. Read
Department of Psychology
Temple University
Philadelphia, PA

Randall T. Salekin
Department of Psychology
University of Alabama
Tuscaloosa, AL

Philip S. Santangelo
Karlsruhe Institute of Technology
Karlsruhe, Germany

Bonnie Solomon
Department of Psychology
University of Illinois at Chicago
Chicago, IL

Lynne Steinberg
Department of Psychology
University of Houston
Houston, TX

David Thissen
Department of Psychology
The University of North Carolina at Chapel Hill
Chapel Hill, NC

Timothy J. Trull
Department of Psychological Sciences
University of Missouri-Columbia
Columbia, MO

Stephen G. West
Department of Psychology
Arizona State University
Tempe, AZ

Emily Wheat
Department of Psychology
Virginia Commonwealth University
Richmond, VA

Erika J. Wolf
National Center for PTSD
VA Boston Healthcare System
Department of Psychiatry
Boston University School of Medicine
Boston, MA

Nina Wong
Anxiety Disorders Clinic
University of Central Florida
Orlando, FL

Michael J. Zvolensky
Department of Psychology
The University of Vermont
Burlington, VT

This page intentionally left blank

CONTENTS

CHAPTER

1 A Place for Research Strategies in Clinical Psychology

Jonathan S. Comer *and* Philip C. Kendall

Abstract

Despite daunting statistics portraying the staggering scope and costs of mental illness, recent years have witnessed considerable advances in our understanding of psychopathology and optimal methods for intervention. However, relative to other sciences, clinical psychology is still a relatively nascent field, and as such the majority of work is ahead of us. The prepared investigator must be familiar with the full portfolio of modern research strategies for the science of clinical psychology. The present Handbook has recruited some of the field's foremost experts to explicate the essential research strategies currently used across the modern clinical psychology landscape. Part I of the Handbook addresses design strategies for clinical psychology and covers laboratory methods in experimental psychopathology, single-case experimental designs, small pilot trials, the randomized controlled trial, adaptive and modular treatment designs, and dissemination methods and models. Part II addresses measurement strategies for clinical psychology and covers assessment, change measurement, observational coding, measurement of process variables across treatment, structural and functional brain imagining, and experience sampling data-collection methods. Part III addresses analytic strategies for clinical psychology and includes chapters on statistical power, correlation and regression, randomized clinical trial data analysis, conventions in mediation and moderation analysis, structural equation modeling, meta-analytic techniques, item-response theory, and the appropriate handling of missing data. In Part IV, matters of responsible conduct in clinical psychology research are covered, including ethical considerations in clinical research and important issues in research with culturally diverse populations. The book concludes with an integrative summary of research strategies addressed across the volume, and guidelines for future directions in research methodology, design, and analysis that will keep our young science moving forward in a manner that maximizes scientific rigor and clinical relevance.

Key Words: Research methods, research strategies, methodology, design, measurement, data analysis

Mental health problems impose a staggering worldwide public health burden. In the United States, for example, roughly half of the population suffers at some point in their lives from a mental disorder (Kessler, Berglund, Demler, Jin, Merikangas, & Walters, 2005), and one in four has suffered from a mental disorder in the past year (Kessler, Chiu, Demler, & Walters, 2005). These estimates are particularly striking when considering the tremendous individual and societal costs associated with mental disorders. When left untreated these disorders are associated with frequent comorbid mental disorders (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Kessler, Chiu, Demler, & Walters, 2005), elevated rates of medical problems (Goodwin, Davidson, & Keyes, 2009; Roy-Byrne, Davidson, Kessler et al., 2006), family dysfunction, disability in major life roles (Merikangas, Ames, Cui, Stang, Ustun, et al., 2007), poorer educational attainment (Breslau, Lane, Sampson, & Kessler,

2008), and overall reduced health-related quality of life (Comer, Blanco, Hasin, Liu, Grant, Turner, & Olfson, 2011; Daly, Trivedi, Wisniewski, et al., 2010). Furthermore, mental disorders confer an increased risk of suicide attempts (Nock & Kessler, 2006) and are prospective predictors of problematic substance use years later (Kendall & Kessler, 2002; Swendsen, Conway, Degenhardt, Glantz, Jin, Merikangas, Sampson, & Kessler, 2010).

The societal burden of mental disorders is portrayed in reports of losses in worker productivity and of high health care utilization and costs (e.g., Greenberg et al., 1999). For example, major depressive disorder (MDD) is associated with workforce impairments, with 20 to 30 percent of Americans with moderate or severe MDD collecting disability and/or unemployed (Birnbaum, Kessler, Kelley, Ben-Hamadi, Joish, & Greenberg, 2010). Depressed workers miss more workdays than nondepressed workers, collectively accounting for roughly 225 million missed annual workdays and corresponding to an estimated $36.6 billion in lost productivity each year (Kessler, Akiskal, Ames, Birnbaum, Greenberg, et al., 2006). Individuals with serious mental health problems earn on average roughly $16,000 less annually than their unaffected counterparts, resulting in estimated total lost annual earnings of $193.2 billion nationally.

Despite these daunting statistics, the past 40 years have witnessed considerable advances in our understanding of psychopathology and the expected trajectories of various disorders, and the field has identified evidence-based interventions with which to treat many of these debilitating conditions (Barlow, 2008; Kendall, 2012). However, much remains to be learned about mental disorders and their treatment, and this should not be surprising. After all, whereas many sciences have been progressing for centuries (e.g., biology, chemistry, physics), it is only recently, relatively speaking, that the scientific method and empiricism have been applied to the field of clinical psychology.

At this relatively early stage in the science of clinical psychology, the majority of work is ahead of us, and as such we must embrace a deep commitment to empiricism and appreciate the intricate interdependence of research and practice as we move forward. The National Institute of Mental Health Strategic Plan (2008) provides a strong guiding framework to focus and accelerate clinical research so that scientific breakthroughs can tangibly improve mental health care and the lives of affected individuals.

First, the past decade has witnessed extraordinary technological advances in our ability to image and analyze the living brain and to collect other biological (e.g., genes, proteins) and experiential data associated with key domains of functioning and dysfunction. Such innovations have the potential to apply noninvasive techniques to understand the development and function of brain networks and how various changes in functional connectivity may place individuals at risk for clinical syndromes and reduced treatment response. Such work can also inform our understanding of neurobiological mechanisms of adaptive and maladaptive change.

Second, despite advances in epidemiology identifying rates and patterns of mental health disorders, longitudinal work is needed to identify developmental patterns of mental disorders in order to determine when, where, and how to intervene optimally. Work in this area evaluating biomarkers may have the potential to identify biosignatures of clinical presentations and treatment response and may help to identify differentially indicated treatments for use at different stages of disorder and recovery. Such work can also help to better identify psychological risk and protective factors across the lifespan.

Third, despite the identification of evidence-based psychological treatment practices with the potential to improve outcomes for many of the mental health problems affecting the population, much remains to be learned to develop interventions for the difficult-to-treat and difficult-to-reach individuals, to improve supported interventions and their delivery, to incorporate the diverse needs and circumstances of affected individuals, and to expand treatment availability, accessibility, and acceptability. Regrettably, substantial problems in the broad availability and quality of psychological treatments in the community constrain effective care for the majority of affected individuals. A new generation of research in clinical psychology is needed to address the gaps that persist between treatment in experimental settings and services available in the community. The blossoming field of dissemination and implementation science (Kendall & Beidas, 2007; McHugh & Barlow, 2010) has begun to systematically address this critical area, but we are just at the very beginning of elucidating optimal methods for broad-based, sustainable training in evidence-based treatments.

Fourth, efforts are needed to expand the public health relevance of clinical research. Innovative research and research strategies are needed that can rapidly inform the delivery of quality treatment to

maximally benefit the largest number of affected or at-risk individuals. Such work would entail comparative-effectiveness analyses and evaluations of supported treatments in nonspecialty settings and by nonspecialty providers across service sectors, while also addressing disparities in care and incorporating technological innovations.

In the face of such grand and laudable objectives for our field, the prepared investigator must be familiar with the full portfolio of modern research strategies for the science of clinical psychology—a set of "directions," so to speak, for getting from "here" to "there." Just as with any travel directions, where many acceptable ways to get to the same destination may exist (e.g., the quick way, the scenic way, the cheap way), for each empirical question there are many research strategies that can be used to reveal meaningful information, each with strengths and limitations. When conducting research, it is incumbent upon the investigator to explicitly know why he or she is taking a particular route, to be familiar with the tradeoffs inherent in taking such a route, and to travel that route correctly.

Importantly, evaluations into psychopathology and therapeutic efficacy and effectiveness have evolved from a historical reliance on simply professional introspection and retrospective case history exploration to the modern reliance on complex multimethod experimental investigations; prospective, longitudinal research; and well-controlled cross-sectional examinations across well-defined samples. The evolution is to be applauded. To continue to move the science of clinical psychology forward, investigators must systematically rely on research strategy "routes" that achieve favorable balances between scientific *rigor* and clinical *relevance*. This requires careful deliberations around matters of tradeoffs between internal validity (which is typically linked with *rigor*) and external validity (which is typically linked with *relevance*). It is with this in mind that we have recruited some of the field's foremost experts for this Handbook to explicate the essential research strategies currently used across the modern clinical psychology landscape that maximize both rigor and relevance.

Part I of the book addresses *design* strategies for clinical psychology and covers laboratory methods in experimental psychopathology, single-case experimental designs, small pilot trials, the randomized controlled trial, adaptive and modular treatment designs, and dissemination methods and models. Part II addresses *measurement* strategies for clinical psychology and covers assessment, change measurement, observational coding, measurement of process variables across treatment, structural and functional brain imagining, and experience sampling data-collection methods.

Part III addresses *analytic* strategies for clinical psychology and includes chapters on statistical power, correlation and regression, randomized clinical trial data analysis, conventions in mediation and moderation analysis, structural equation modeling, meta-analytic techniques, item-response theory, and the appropriate handling of missing data. In Part IV, matters of *responsible conduct* in clinical psychology research are covered, including ethical considerations in clinical research and important issues in research with culturally diverse populations. The book concludes with an integrative summary of research strategies addressed across the volume, and guidelines for future directions in research methodology, design, and analysis that will keep our young science moving forward in a manner that maximizes scientific rigor and clinical relevance.

References

Barlow, D. H. (Ed.). (2008). *Clinical handbook of psychological disorders* (4th ed.). New York : Guilford Press.

Birnbaum, H. G., Kessler, R. C., Kelley, D., Ben-Hamadi, R., Joish, V. N., & Greenberg, P. E. (2010). Employer burden of mild, moderate, and severe major depressive disorder: Mental health services utilization and costs, and work performance. *Depression and Anxiety*, *27*(1), 78–89. doi:10.1002/da.20580

Breslau, J., Lane, M., Sampson, N., & Kessler, R. C. (2008). Mental disorders and subsequent educational attainment in a US national sample. *Journal of Psychiatric Research*, *42*(9), 708–716.

Comer, J. S., Blanco, C., Hasin, D. S., Liu, S. M., Grant, B. F., Turner, J. B., & Olfson, M. (2011). Health-related quality of life across the anxiety disorders: Results from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *Journal of Clinical Psychiatry*, *72*(1), 43–50.

Costello, E., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*(8), 837–844. doi:10.1001/archpsyc.60.8.837

Daly, E. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Gaynes, B. N., Warden, D., &…Rush, A. (2010). Health-related quality of life in depression: A STAR*D report. *Annals of Clinical Psychiatry*, *22*(1), 43–55.

Goodwin, R. D., Davidson, K. W., & Keyes, K. (2009). Mental disorders and cardiovascular disease among adults in the United States. *Journal of Psychiatric Research*, *43*(3), 239–246. doi:10.1016/j.jpsychires.2008.05.006

Greenberg, P. E., Sisitsky, T., Kessler, R. C., Finkelstein, S. N., Berndt, E. R., Davidson, J. R., et al. (1999). The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry*, *60*, 427–435.

Kendall, P. C. (2012). *Child and adolescent therapy: Cognitive-behavioral procedures* (4th ed.). New York : Guilford.

Kendall, P. C., & Beidas, R. S. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology: Research and Practice*, *38*, 13–20.

Kendall, P. C., & Kessler, R. C. (2002). The impact of childhood psychopathology interventions on subsequent substance abuse: policy implications, comments, and recommendations. *Journal of Consulting and Clinical Psychology*, *70*(6), 1303–1306.

Kessler, R. C., Akiskal, H. S., Ames, M., Birnbaum, H., Greenberg, P., Hirschfeld, R. A., &… Wang, P. S. (2006). Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers. *American Journal of Psychiatry*, *163*(9), 1561–1568. doi:10.1176/appi.ajp.163.9.1561

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *62*(6), 593–602. doi:10.1001/archpsyc.62.6.59

Kessler, R. C., Chiu, W., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *62*(6), 617–627. doi:10.1001/archpsyc.62.6.617

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist*, *65*, 73–84.

Merikangas, K. R., Ames, M., Cui, L., Stang, P. E., Ustun, T., Von Korff, M., & Kessler, R. C. (2007). The impact of comorbidity of mental and physical conditions on role disability in the US adult household population. *Archives of General Psychiatry*, *64*(10), 1180–1188. doi:10.1001/archpsyc.64.10.1180

National Institute of Mental Health (2008). *National Institute of Mental Health Strategic Plan.* Bethesda, MD : National Institute of Mental Health.

Nock, M. K., & Kessler, R. C. (2006). Prevalence of and risk factors for suicide attempts versus suicide gestures: Analysis of the National Comorbidity Survey. *Journal of Abnormal Psychology*, *115*(3), 616–623.

Roy-Byrne, P. P., Davidson, K. W., Kessler, R. C., Asmundson, G. G., Goodwin, R. D., Kubzansky, L., &… Stein, M. B. (2008). Anxiety disorders and comorbid medical illness. *General Hospital Psychiatry*, *30*(3), 208–225. doi:10.1016/j.genhosppsych.2007.12.006

Swendsen, J., Conway, K. P., Degenhardt, L., Glantz, M., Jin, R., Merikangas, K. R., &… Kessler, R. C. (2010). Mental disorders as risk factors for substance use, abuse and dependence: Results from the 10-year follow-up of the National Comorbidity Survey. *Addiction*, *105*(6), 1117–1128. doi:10.1111/j.1360–0443.2010.02902.x

PART 1

Design Strategies for Clinical Psychology

This page intentionally left blank

CHAPTER

2 Laboratory Methods in Experimental Psychopathology

Michael J. Zvolensky, John P. Forsyth, *and* Kirsten Johnson

Abstract

Experimental psychopathology represents a subfield of psychological science aimed at elucidating the processes underlying abnormal behavior. The present chapter provides a synopsis of key elements of experimental psychopathology research and its methods. In the first section, we define experimental psychopathology research and briefly articulate its origins. Second, we present the methodological approaches employed in experimental psychopathology research. Third, we present some of the molar conceptual considerations for the assessment approaches in experimental psychopathology research. In the final section, we describe some key challenges to experimental psychopathology research as well as potentially useful strategies recommended for overcoming such challenges.

Key Words: Experimental psychopathology, laboratory, mechanism, laboratory models, translational

Experimental psychopathology represents a subfield of psychological science aimed at elucidating the processes underlying abnormal behavior (Zvolensky, Lejuez, Stuart, & Curtin, 2001). Although originally restricted to "true experimental" laboratory-based tests (Kimmel, 1971), experimental psychopathology now reflects a larger, more diverse and multifacted field of inquiry (Zvolensky et al., 2001). Topics of study include examinations of the phenomenology of psychological disorders; explication of the underlying processes governing the etiology, maintenance, and amelioration of psychopathology; and tests of intervention(s) with the express purpose of identifying explanatory mechanisms. This work typically involves a range of methodologies (e.g., laboratory and field studies) as well as populations (e.g., diagnosed cases and nonclinical). The subfield of experimental psychopathology represents one of the branches in psychological science upon which evidence-based practice is theoretically and empirically built (Forsyth & Zvolensky, 2002; McFall, 1991).

The present chapter provides a synopsis of the key elements of experimental psychopathology research and its methods. In the first section, we define experimental psychopathology research and briefly articulate its origins. Second, we present the methodological approaches employed in experimental psychopathology research. Third, we present some of the molar conceptual considerations for the assessment approaches in experimental psychopathology research. In the final section, we describe some key challenges to experimental psychopathology research as well as potentially useful strategies recommended for overcoming such challenges.

Experimental Psychopathology: Definition and Origins

Definition

Kimmel (1971) offered one of the earliest definitions of experimental psychopathology research: "*the experimental study of pathological behavior* (i.e., using the experimental method to study pre-existing pathological behavior), or *the study of experimental*

pathological behavior (i.e., the pathological behavior being studied is induced experimentally rather than developed naturally)" (p. 7, emphasis added). In the former sense, experimental psychopathology is the study of the behavior of individuals with known psychopathology in response to imposed experimental conditions (e.g., how persons with and without a diagnosis of a specific disorder respond under conditions of "imposed stress" or "no stress"), whereas the latter approach entails identification and manipulation of variables to induce psychopathology processes among individuals without a history of psychopathology (Forsyth & Zvolensky, 2002).

Others have defined experimental psychopathology more broadly as the application of methods and principles of psychological science to understand the nature and origins of abnormal behavior (Kihlstrom & McGlynn, 1991). This definition encompasses other quasi-experimental and correlational methodologies (Kihlstrom & McGlynn, 1991). Zvolensky and colleagues (2001) defined experimental psychopathology as laboratory-based research with human and/or nonhuman animals, directly aimed at discovering and/or explaining the etiology and maintenance of psychopathological processes; work that may potentially contribute to the amelioration of dysfunctional behavior in the future. These definitions of experimental psychopathology can be contrasted with that of clinical psychopathology research that involves study with humans, typically with a particular psychological disorder, to (a) address the treatment/prevention of psychopathology in settings primarily outside of the laboratory or (b) identify symptoms or deficits that characterize psychological disorders (Forsyth & Zvolensky, 2002). Moreover, experimental psychopathology can be distinguished from "basic psychological research." Although basic research ultimately may have important clinical implications, the goals and overarching framework for such research are to elucidate broadly applicable principles, independent of any *a priori* clinical relevance or application (Osgood, 1953).

Across perspectives, the common thread that runs through each of the above definitions of experimental psychopathology is a focus on knowledge development for psychopathology by using experimental and related methodologies. The phenomenon of interest can be induced or it may consist of already occurring natural abnormal behavior. Please see below ("Underlying Conceptual Approach") for an expanded discussion of the main methodological approaches employed in experimental psychopathology research.

Origin

The origin of experimental psychopathology can be discussed in relation to the scholarly work of Ivan Pavlov (1849–1936) and William James (1842–1910). Both Pavlov and James helped to establish two traditions within experimental psychopathology: (a) the experimental induction and subsequent modeling of abnormal behavior in laboratory animals (Pavlov) and (b) the experimental approach to the study of preexisting abnormal behavior in humans (James).

Pavlov first used the label "experimental psychopathology" in a 1903 lecture delivered at the International Medical Congress in Madrid titled *Experimental Psychology and Psychopathology in Animals* (Forsyth & Zvolensky, 2002). In that lecture, Pavlov presented for the first time his theory of conditioned reflexes, which revealed his tendency to cast psychopathology in physiological terms via experimental tests on animals. Yet Pavlov did not develop a coherent conceptualization of experimental psychopathology apart from use of an "experimental approach." Later, the contributions of two different investigators in Pavlov's laboratory facilitated advancements in the area of experimental psychopathology research (Kimmel, 1971; Popplestone & McPherson, 1984). Specifically, Yerofeeva (1912, 1916) and Shenger-Krestovnikova (1921) both observed persistent "abnormal behavior" in their experimental animals following the use of novel conditioning procedures. This work was the precursor to the phenomenon later known as "experimental neurosis." This work on experimental neurosis led to a marked shift in Pavlov's research agenda: he devoted the remainder of his scientific career to the experimental analysis of variables and processes that occur in abnormal behavior patterns among human and nonhuman animals.

Due to Pavlov's contributions, other behavioral scientists began to view laboratory approaches to studying abnormal behavior processes as both meaningful and productive (Benjamin, 2000). From this work emerged a core conceptual principle of subsequent experimental psychopathology research. That is, otherwise adaptive behavioral processes provide the foundation upon which maladaptive patterns of behavior are established, and such behavioral processes can subsequently interfere with an organism's ability to behave effectively (Kimmel, 1971). The core task, therefore, involved isolating

processes responsible for "moving" the organism from an adaptive to a maladaptive range of behavior. Notably, although not the present purpose, this core experimental psychopathology concept helped pave the way for comparative psychiatry (Lubin, 1943) and comparative psychopathology (Zubin & Hunt, 1967)—approaches that emphasize cross-species comparisons of behavioral processes (e.g., human-to-animal comparisons).

This approach of focusing on identifying and manipulating variables that, either in whole or in part, cause and/or exacerbate psychopathology began to define experimental psychopathology research (Kimmel, 1971). Yet while Pavlov and his contemporaries confined themselves to the experimental production of psychopathological behavior in the laboratory (Anderson & Liddell, 1935; Krasnogorski, 1925; Liddell, 1938), William James had been working to develop an experimental psychopathology of abnormal behavior as it occurs naturally.

Although William James is best known for his functionalist philosophy, he also helped to pioneer work in experimental psychopathology. James was highly critical of trends in American psychology; indeed, he was particularly judgmental of the importing of the German ideal of science (i.e., Wundtian and Titchenerian psychology), with its emphasis on determinism, materialism, structuralism, and reductionism (Taylor, 1996). James believed the Wundtian experimental tradition had created a less-than-ideal instrument, a tradition characterized by laboratory activities focused on building apparatus to collect "trivial measurements." Specifically, this work often lacked practical purpose or relevance (Taylor, 1996). James was chiefly concerned that psychology might lose sight of developing a comprehensive and integrated psychological science of the "whole person."

William James established a laboratory at Harvard University in 1875. Between 1874 and 1889, James was involved in collaborative research in Bowditch's laboratory at Harvard Medical School (Taylor, 1996). While at Harvard University, James maintained active collaborations with individuals at Harvard Medical School in an attempt to bridge areas of physiology, neurology, and psychology, an approach well ahead of its time (National Advisory Mental Health Council Behavioral Science Workgroup, 2000). As Taylor (1996) observed, these laboratory sciences became integrated at Harvard, culminating in experimental research on the problem of consciousness. James, in particular, addressed the problem of consciousness via experiments on hypnosis, automatic writing, phantom limb phenomenon, psychophysical manipulations with clinical samples (e.g., perception of space, balance), neurophysiology, and studies of dissociative phenomena. These topics of study and the methods used to examine them represent the precursors to a science of abnormal behavior, and experimental psychopathology specifically.

According to James, experimental psychopathology was the study of the variables and processes that influence both aberrant and exceptional human experience. James's laboratory work was largely devoted to an experimental analysis of psychopathology as it occurs naturally. James also conducted his experimental psychopathology research with a focus on practical concerns, a direct precursor to topics that now carry the label "clinical relevance." This approach also was heavily influenced by the clinical emphasis of the emerging French laboratory sciences in physiology, neurology, experimental physiology, and psychology. Notably, this approach can be contrasted with the Germanic tradition, where pure science was considered to be the father of clinical application (Taylor, 1996). The infusion of ideas emerging from rapid developments in experimental psychopathology gave way to the rise of scientific psychotherapy in America at a time when experimental psychology, psychiatry, and medicine also were beginning to question the practices of mental healers (Taylor, 1996); this move is remarkably similar to contemporary efforts to stymie "pseudoscience" activities (Lilienfeld, 1996).

Pavlovian- and Jamesian-style experimental psychopathology research gained momentum through the early to middle nineteenth century. By 1904, the first world congress of experimental psychology was convened (Schumann, 1905). By 1910, several texts appeared outlining current experimental psychopathology research, most notably Gregor's (1910) *Leitfaden der Experimentellen Psychopathologie* ("guide" or "how-to book" for experimental psychopathology). Two years later, Franz (1912) published the first review of experimental psychopathology research in *Psychological Bulletin*; a review that was followed 4 years later by another review article with the same title (Wells, 1914). Neither Franz nor Wells offered a definition of experimental psychopathology in their respective works, yet both papers are of historical interest in highlighting the nature of experimental psychopathology research during this period. Much of this research, in turn, was occurring in the context of various labels, such as

abnormal psychology, psychopathology, pathopsychology, clinical psychology, psycho-clinical, medical psychology, and medico-clinical. Moreover, this work was experimental in design and focused on questions pertaining to the understanding of normal and abnormal psychological processes.

Experimental psychopathology research proliferated in the ensuing decades and tended to follow either a Jamesian (i.e., experimental analysis of naturally occurring psychopathology) or Pavlovian (i.e., the experimental induction of psychopathology) approach. The behaviorists were following the lead of Pavlov, Hull, and Watson in pursuing basic and applied work in experimental neurosis (e.g., Estes & Skinner, 1941; Franks, 1964; Rachman, 1966; Skinner, 1953; Wolpe, 1952, 1958; Wolpe, Salter, & Reyna, 1964). This approach drew heavily upon the findings from experimental psychology and involved a conception of abnormality in terms of deficits in functioning of certain psychological systems rather than people suffering from mental diseases produced by biological causes (Eysenck, 1973). Notably, Eriksonians, Gestaltists, and Freudians also were embarking on experimental psychopathology and outlining a framework for how such work might proceed (e.g., Brown, 1937; Mackinnon & Henle, 1948). Additionally, numerous attempts were under way to extend findings of experimental psychopathology to the practice of psychotherapy (Landis, 1949; Masserman, 1943; Wolpe, 1958) and to use this work as a framework for a science of abnormal behavior, generally (Eysenck, 1961). By 1947, the American Psychological Association (APA) recognized experimental psychopathology research as a legitimate component of accredited training in clinical psychology (APA, 1947).

Experimental psychopathology research grew further in the early 1950s with the establishment of the *Journal of Clinical and Experimental Psychopathology*. This journal became an outlet for experimental psychopathology research. The 1955 APA meeting also was significant in its thematic focus on the experimental approach to psychopathology (see Hoch & Zubin, 1957). Experimental psychopathology grew and diversified during this period, and by the early to middle 1960s began to include laboratory observation of psychopathological processes. Indeed, the 1960s marked an important historical shift in focus and a more widespread usage of the word "experimental" to include research, often in a laboratory setting, but where the purpose was to identity psychopathological processes. Since this period, numerous professional organizations and journals have been developed that showcase and disseminate experimental psychopathology research. For example, the *Journal of Abnormal Psychology* is a flagship journal of the APA that has played a key role in experimental psychopathology dissemination.

Unlike "true experiments," where the focus is to vary some variable deliberately after random assignment so as to elucidate causal relations, the new wave of experimental psychopathology often utilized correlational methodology (Kimmel, 1971). This approach now often falls under the label of "descriptive psychopathology research." The global purpose of descriptive psychopathology research is to identify markers that are thought to characterize, or covary with, psychopathological processes of phenotypes. Although markers can be either broadband or specific to forms of psychopathology, the concept itself includes examination of individual difference variables thought to aid in the prediction, diagnosis, or understanding of the consequences of a disorder (Sher & Trull, 1996). Such markers are typically studied via use of sophisticated laboratory methods that may involve biochemical assays, pharmacological or psychological challenges, psychophysiological measures, neuropsychological assessments, or cognitive assessments (see Hunt & Cofer, 1944; Kihlstrom & McGlynn, 1991; Lenzenweger & Dworkin, 1998; Sher & Trull, 1996, for more comprehensive descriptions of this approach).

The move from a strict application of experimental to descriptive psychopathology research methodology appears to be greatly influenced by advances in cognitive and neuropsychological assessment instruments and research. Here, the focus often is to understand higher-order cognitive processes of relevance to various psychopathological states (e.g., executive functioning, language abilities, and attentional functions) via sophisticated instruments. This work also appears to have been driven by early refinements in the *Diagnostic and Statistical Manual of Mental Disorders* (e.g., DSM; American Psychiatric Association, 1994). This shift in focus not only contributed to understanding the nature of abnormal behavior (Chapman & Chapman, 1973; Ingram, 1986; McNally, 1998), but it also provided important insights into the role of cognitive functioning in the development, expression, and maintenance of psychopathology (Abramson & Seligman, 1977; Kihlstrom & McGlynn, 1991; Maser & Seligman, 1977).

Overall, experimental psychopathology research has grown from basic laboratory roots and in many respects represents a hybrid of other laboratory

and clinical research. It has been influenced by a number of philosophical, contextual, and methodological developments over the past 100-plus years. Currently, experimental psychopathology tends to reflect work that is concerned with underlying processes for psychopathology and often examines them via experimental or correlational methodology. With this background, we now turn to a more in-depth discussion of the main methodological approaches employed in experimental psychopathology research.

Underlying Conceptual Approach

Forsyth and Zvolensky (2002) derived a two-dimensional scheme for characterizing experimental psychopathology work. The first dimension spans research where an independent variable is manipulated or not manipulated. The second dimension includes the nature of the population under study. The resulting matrix yields four possible ways to characterize experimental psychopathology research (i.e., Type I, Type II, Type III, and Type IV). Please see Table 2.1 for a summary of these labels and their definitions.

Type I Experimental Psychopathology Research

Type I research involves the manipulation of independent variables and examination of their effects on behavior in nonclinical samples. Although both dimensions characterize experimental psychology research, they represent experimental psychopathology research when the *a priori* focus is on elucidating processes that contribute, either in whole or in part, to the genesis or maintenance of abnormal behavior (see Table 2.1). In short, these studies can address "if," "how," and "why" questions concerning pathogenic variables and processes. Included here would be studies that attempt to produce critical features of psychopathology in organisms with no known history of psychopathology (e.g., experimental neurosis; Pavlov, 1961; Wolpe, 1958). In such work, psychopathology processes represent *the* dependent variable of interest. These processes are often induced directly in mild but prototypical forms.

Given the focus on the induction of psychopathology, participants in Type I research often are those who have no preexisting history of psychopathology. Such participants, unlike those with preexisting psychopathology, provide experimental psychopathologists with a relatively "clean" biobehavioral history upon which to engage in theory-driven causal-oriented hypothesis testing. Moreover, bodies of work on a particular type of process (e.g., respondent learning during physical stress) theoretically offer a "normative context" upon which sophisticated evaluations of similar processes in clinical samples can be better understood. Although it is sometimes common to challenge the use of nonclinical samples in these studies, such arguments are often not theoretically justified from an experimental psychopathology research tradition. Indeed, experimental psychopathology seeks to determine, on an *a priori* basis, the nature of specific biobehavioral processes moving one from a normal psychological experience to a dysfunctional one. As indicated, the basic assumption guiding this work is that abnormal behavior is governed by the same principles and classes of variables that determine normal adaptive behavior. It is the combination of

Table 2.1. Classifications and Definitions of Psychopathology Research

Dimension	Definition
Type I: Experimental psychopathology research	The manipulation of independent variables to observe their effects on behavior in nonclinical samples. Here, the *a priori* focus is on elucidating variables that contribute to the genesis of abnormal behavior.
Type II: Quasi-experimental psychopathology research	The manipulation of independent variables to observe their biobehavioral processes in samples with a well-established type of psychopathology, among persons displaying well-established or subclinical features of psychopathology
Type III: Nonpatient psychopathology research	No manipulation of independent variables; is limited to descriptive statements about behavioral and psychological processes in nonclinical samples
Type IV: Descriptive psychopathology research	No manipulation of independent variables; is limited to descriptive statements about psychopathology in samples with well-established or subclinical features of psychopathology

such variables that results in variations in behavior, some of which may be characterized as abnormal or maladaptive in a psychological sense (Sandler & Davidson, 1971). The task, therefore, is to examine how varied permutations of such variables result in psychopathological behavior. Thus, such questions are impossible to address among those already experiencing psychopathology. For example, Peters, Constans, and Mathews (2011) employed a Type I research paradigm to test the hypothesis that attributional style may be one causative factor of depression vulnerability. Here, 54 undergraduate students, without a history of depression, were randomly assigned to one of two experimental conditions: resilience condition, $n = 28$; vulnerability condition, $n = 26$. The resiliency condition involved exposing participants to 60 descriptions that promoted a self-worthy, stable attribution of a positive event and 60 descriptions that promoted an unstable attribution unrelated to self-worth for a negative event. In contrast, the vulnerability condition involved exposing participants to 60 descriptions that promoted a self-deficient, stable attribution of a negative event and 60 descriptions that promoted an unstable attribution unrelated to self-worth for a positive event. Following exposure to the assigned descriptions, all participants subsequently completed a stressor task (i.e., Cognitive Ability Test). Through a series of assessments, Peters and colleagues measured the change in mood state from before to after manipulation. Results indicated that individuals in the resilience condition reported less depressed mood (compared to the vulnerability condition) in response to the academic stressor (please see Peters et al., 2011, for a complete description of study methods and results).

Notably, Type I models naturally do not yield a complete account of how psychopathology develops. The reason is that experimental psychopathology models tend to be highly local and specific, and for ethical, pragmatic, and strategic reasons also tend to focus on specific subsets of variables in relation to the induction of prototypical aspects of abnormal behavior. That is, the variables shown to produce key features of psychopathology in a specified population represent only a subset of a universe of possible causal variables that may be subjected to experimental scrutiny in the relatively closed system of the laboratory. Although it often is assumed that such variables will lead to similar behavioral effects in the open system of the natural world (Mook, 1983), this may not always be true. The open system is subject to many dynamic influences, and psychopathology within such a system is often a product of multiple controlling variables. Type I models, therefore, should be viewed not as exact replicas of psychological disorders, or as *the* model of a specific form of psychopathology based on correspondence alone.

The logic of Type I experimental psychopathology research is similar to that of basic research: to yield scientific understanding of the independent variables and relevant processes that cause or maintain forms of psychopathology. This approach is guided by the view that diagnostically similar and dissimilar forms of psychopathology (American Psychiatric Association, 1994) are complexly determined. Thus, two individuals who meet identical DSM diagnostic criteria for Disorder X may exhibit markedly different histories with respect to causal variables (i.e., equifinality), just as two individuals who meet criteria for different DSM diagnoses may exhibit fairly similar histories with respect to putative causal variables (i.e., multifinality). The task of Type I experimental psychopathology research, therefore, is to elucidate a universe of relevant causal processes and their relation to specific forms of psychopathology.

Such Type I research is not driven by, nor necessarily dependent on, a reliable and valid psychiatric nomenclature (but see Kihlstrom & McGlynn, 1991, for a different view). Indeed, Type I experimental psychopathology may be inspired by the psychiatric nomenclature, or more generally by questions about the nature of psychological suffering and behavior maladjustment (e.g., Gantt, 1971). The expectation over time is that this work will yield a clearer understanding of a subset of clinically relevant variables.

Type I research has a strength (and challenge) of being able to identify putative causal variables that are directly manipulable. This aspect is important for this research approach, as such variables, to the extent that they are subject to direct manipulation, also may serve as the "building blocks" of future intervention efforts. Despite the apparent analytic correspondence that is involved with the identification of "controlling variables" and the direct application of such variables to intervention strategies, it is indeed rare that experimental psychopathologists follow these processes fully through to the point of application (see Zvolensky et al., 2001, for a discussion of this issue). The reason is due, in part, to the analytic agenda of Type I experimental psychopathologists. This agenda is constrained by analytic goals of prediction and influence, and the more

general epistemic agenda of contributing to scientific understanding (i.e., knowledge for knowledge's sake), and only secondary concern about whether such knowledge may be put to practical use.

Type II Experimental Psychopathology Research

Type II research involves the direct manipulation of independent variables and evaluation of their effects on biobehavioral processes in samples with a well-established type of psychopathology, among persons who vary in some established psychopathology risk dimension or display subclinical (i.e., they do not reach a diagnostic threshold) features of psychopathology. For example, Felmingham and colleagues (2010) recorded functional magnetic resonance imaging data in both male and female participants with a diagnosis of posttraumatic stress disorder (PTSD), trauma-exposed controls, and non–trauma-exposed controls while they viewed masked facial expressions of fear. Specifically, fear and neutral gray-scale face stimuli were presented in a backward masking paradigm, with target stimuli (fear or neutral) presented for 16.7 ms, followed immediately by a neutral face mask (163.3 ms). By examining neural activation to threat, Felmingham and colleagues sought to elucidate one of the possible pathways through which women have a greater propensity than men to develop PTSD following trauma. Findings indicated that exposure to trauma was associated with enhanced brainstem activity to fear in women, regardless of the presence of PTSD; however, in men, brainstem activity was associated only with the development of PTSD. Moreover, men with PTSD (compared to women) displayed greater hippocampal activity to fear, possibly suggesting that men have an enhanced capacity for contextualizing fear-related stimuli (please see Felmingham et al., 2010, for a complete description of study methods and results). As illustrated here, unlike Type I experimental psychopathology research, Type II research is limited to quasi-experimental questions of the "what," "if," and "how" variety. Type II research cannot directly provide answers to "why" questions because the psychopathology is selected for, not produced directly. Although this type of research can attempt to address questions about variables and processes that "cause" psychopathology, it is unable to do so in a clear and unambiguous sense. The central reason for this analytic limitation boils down to this: because the variables responsible for a given psychopathology are unknown (at least in part), one cannot clearly demonstrate the effects of independent variables on the naturally occurring psychopathology; thus, it cannot be clearly shown that the independent variables are related to the psychopathology in a causative sense. Please refer to Table 2.1.

Behavior characterized at one point in time as "abnormal" is presumably the product of complex interactions of causative variables and biobehavioral processes associated with them. A psychiatric diagnosis is a summary label of that cumulative history but is not synonymous with it. That is, although one can assume that psychopathology is the result of a history of causative and controlling variables that are somehow different from persons who do not meet diagnostic criteria, one cannot infer from the diagnosis the putative variables and processes responsible for it. Thus, when independent variables are varied among persons with Diagnosis X, any resulting changes in behavior may be due to the interaction of the independent variable and a host of unknown variables and processes in persons so diagnosed. The result leaves the experimenter hypothesizing about why the changed independent variable functioned differentially in one patient sample compared to another. For instance, research has shown that persons with a diagnosis of panic disorder are more likely to panic in response to biological challenges than persons with other anxiety disorder diagnoses and healthy nonpsychiatric controls (Zvolensky & Eifert, 2000). What remains entirely unclear from this research, however, is *why* persons with a diagnosis of panic disorder are more likely to panic in response to biological challenges. As the vast majority of psychological and pharmacological treatment strategies are geared toward implementing a treatment based upon—almost exclusively—psychiatric diagnosis, the "why" question has at first glance very little practical value. To be sure, there is a real and powerful temptation to attribute the cause of differential responses to biological challenge procedures to a psychiatric diagnosis. In doing so, however, the variables responsible for the differential responses are left unexplained.

From a scientific standpoint, the biobehavioral processes associated with a diagnosis of panic disorder or other psychological disorders, and particularly their interaction with a challenge procedure or other experimental manipulations, cannot be fully addressed with Type II research. Although Type II research is specific to descriptive (correlational) statements and its relation to other processes, this need not always be the case. For instance, Type I and Type II experimental psychopathology research

can be programmatically combined, such that the "psychopathology" is experimentally induced (i.e., Type I) and then subjected to other experimental conditions (i.e., Type II). In this way, one can move closer to addressing how variables responsible for producing the psychopathology interact with variables that may either exacerbate or attenuate a range of behavior associated with psychopathology.

Overall, Type II research can elucidate independent variables that (a) exacerbate, or modify the expression of, existing forms of abnormal behavior (e.g., pathoplastic effects; Clark, Watson, & Mineka, 1994) and (b) may be influenced directly to either prevent or ameliorate psychopathology (e.g., treatment intervention as an independent variable). This work occupies an important place in the broader scientific context. For instance, pressing clinical concerns often focus on mechanisms that may be prototypical "gateways" for other types of destructive or problematic behaviors. Similarly, psychopathologists may attempt to elucidate how the presence or absence of certain variables or conditions either increases or decreases the risk for a specific type of behavior, including how such variables may exacerbate the clinical severity of a preexisting psychological condition. Yet when Type II research focuses on testing the efficacy of manipulable treatment interventions on therapeutic outcomes, it more likely belongs within the realm of clinical psychopathology research (see Kihlstrom & McGlynn, 1991; Sher & Trull, 1996). Finally, Type II experimental psychopathology is dependent on the reliability and validity of the psychiatric nomenclature, including methods used to identify and discriminate persons with known or subclinical forms of psychopathology from "normals," based largely on topographic or symptom features alone.

Type III "Nonpatient" Psychopathology Research

Unlike research of the Type I and Type II varieties, Type III research involves no manipulation of independent variables and is limited to descriptive statements (i.e., largely correlational) about behavioral and psychological processes in nonclinical samples. For example, Proctor, Maltby, and Linley (2011) recruited 135 nonclinical undergraduate students to complete self-reported measures of strengths use, subjective well-being, self-esteem, self-efficacy, health-related quality of life, and values-in-action. Here, Proctor and colleagues generated descriptive statements regarding the most- and least-commonly endorsed character strengths as well as the relation(s) observed between these strengths and overall life satisfaction. Specifically, results indicated that the values-in-action strengths of hope and zest were significant positive predictors of life satisfaction (please see Proctor et al., 2011, for a complete description of study methods and results). Although some Type III research could, in principle, contribute to understanding psychopathology (e.g., elucidating behavioral or individual difference risk factors associated with psychological problems), often the goals of such research are not specific to questions about abnormal behavior per se. Only when Type III research is embedded within the larger context of clinical science and practice may it become relevant to understanding psychopathology; this topic is beyond the scope of the present paper. Please refer to Table 2.1.

Type IV Descriptive Psychopathology Research

Type IV research involves no manipulation of independent variables and is thus limited to either descriptive or correlational statements about psychopathology in samples with known or subclinical features of psychopathology. Please refer to Table 2.1. As with Type II, the nature of the population under study (i.e., clinical or subclinical individuals) most clearly identifies Type IV research as belonging within the realm of psychopathology research. As such, Type IV research also is predicated on the reliability and validity of the DSM diagnostic system, including related methods of classification. Type IV research has become increasingly popular in recent years, owing much to the growing precision of psychiatric diagnosis and interest in delimiting characteristic features of different forms of abnormal behavior. This work, in turn, draws heavily on sophisticated assessment methodologies and tasks, many of which are drawn from experimental psychology and medical research (Kihlstrom & McGlynn, 1991). For example, Hawkins and Cougle (2011) examined the relation(s) between anger and a variety of clinically diagnosed anxiety disorders among participants in a large, nationally representative survey. Using a combination of self-report measures and structured clinical interviews, Hawkins and Cougle provided correlational statements about the possible link between anxiety-related psychopathology and the expression of anger. Specifically, results of this investigation suggest that there are unique relationships between multiple anxiety disorders (excluding panic disorder and PTSD) and various indices of anger experience and expression that are not better

accounted for by psychiatric comorbidity (please see Hawkins & Cougle, 2011, for a complete description of study methods and results).

Use of such "experimental" tasks in the context of Type IV research sometimes can give the impression that such research is experimental. Yet use of an experimental task does not *ipso facto* entail that the research is experimental, and hence, capable of addressing questions about variables and processes that maintain, exacerbate, or attenuate psychopathology. Type IV research usually includes the application (not manipulation) of experimental tasks in the context of elucidating biobehavioral differences between clinical and nonclinical samples (e.g., see Kihlstrom & McGlynn, 1991; McNally, 1998; Williams, Mathews, & MacLeod, 1997). Typically, any observed differences are then used to support inferences about the nature of the psychopathology in question, including presumed underlying dysfunctional processes thought to covary with known forms of psychopathology. Much of this work can be classified as descriptive or demonstration psychopathology studies. This is a direct acknowledgment that Type IV research can inform our understanding about what persons with different forms of abnormal behavior typically do in response to imposed tasks under controlled conditions, but not why they do what they do.

Summary

Four common types of experimental psychopathology research differ in their focus on manipulation of independent variables and sample type (Forsyth & Zvolensky, 2002). These types of research vary in their ability to identify processes governing the origins and maintenance of psychopathology processes. Yet across these types of research there are some overarching assessment issues that are routinely considered. We now turn to a discussion of these molar conceptual considerations in the context of experimental psychopathology research.

Assessment Approach in Experimental Psychopathology Research: Molar Conceptual Considerations

The assessment approach for experimental psychopathology research has no single "strategy" that will work for all types of scientific activities. There also is no standard model that can work for all types of experimental psychopathology research. Yet a number of basic issues, including level of analysis, method of assessment, nature of inferences drawn, and quality of the data obtained, provide a conceptual basis for understanding how and why certain assessment activities are employed in any given type of experimental psychopathology research.

Level of Analysis

In most instances, the procedures employed to execute assessment activities in experimental psychopathology research are highly influenced by the underlying conceptual framework for the psychopathology phenotype in question (Kazdin, 1982). For example, the level of analysis for the assessment of psychopathology processes is largely influenced by the conceptualization of the problem behavior in question. In most cases, assessment activities for experimental psychopathology focus on symptom presentation (e.g., number of panic attacks per observational period), psychopathology phenotype (e.g., alcohol abuse vs. dependence), or the operative system components (cognitive, behavioral, physical, and social context). The level of analysis employed in experimental psychopathology will directly affect the extent to which specific aspects of problematic behavior are assessed.

Assessment at the symptom level in experimental psychopathology focuses on individual behavior (e.g., number of drinks per drinking episode, number or intensity of catastrophic thoughts); it is a unidimensional approach. Assessment at the phenotypic level focuses on the symptoms that covary, and therefore it is multidimensional (e.g., facets of distinct elements of drinking behavior or thought processes); this approach encompasses more elements of the individual's behavior (e.g., frequency, amount, consequences). Assessment at the system level tends to be more inclusive, assuming that the various systems involved affect one another in a direct fashion; for example, substance use behavior affects anxiety and related mood states and vice versa (Drasgow & Kanfer, 1985). Although more inclusive theoretically, the challenge to using the system-level approach historically has been in the titration of the accuracy operative conceptual model in terms of the pragmatic aspects of the assessment processes (e.g., isolating the appropriate level to assess problem behavior relative to existing scientific information about it).

Methods

All levels of analysis for the assessment of experimental psychopathology can theoretically involve the measurement of responses across cognitive, behavioral, and physiological systems. The measurement of specific systems varies both by content area

(e.g., depressive vs. anxiety disorder) and the particular systems theoretically involved with the problem behavior in question. Therefore, there is great variability across distinct types of psychopathology despite recognition of some of their overarching commonalities. The classic work by Cone (1978) provides a model for understanding the assessment process in experimental psychopathology research. Cone (1978) identified that assessment tactics vary along dimensions—content, directness, and generalizability. Content reflects the nature of the responses being assessed (cognitive, behavioral, and physiological). Directness pertains to the immediacy of the assessment of responses in the time and context in which they occur (e.g., measuring alcohol use during periods of actual use vs. retrospective reports of alcohol use behavior). Common forms of indirect methods of assessment include interviews, questionnaires, and ratings by self or others; common forms of direct assessment include monitoring behavior in real-world settings (e.g., time sampling approaches), role playing, and various forms of analogue behavior (e.g., measuring emotional responses to drug cues in the laboratory). Generalizability refers to the consistency of the responses being measured across a particular domain. There are distinct domains of generalizability often relevant to psychopathological processes (e.g., time, setting, method; Cone, 1978).

Contingent upon the goals of the assessment, there will be natural variation in the method and content targeted for measurement in experimental psychopathology. There also are likely differences in method and content during assessment as a function of the training and background of the assessor. There is naturally no "universal" or "correct" model that will be sufficient to meet the assessment objectives for all types of psychopathology. In short, the methods employed to assess the content areas of primary interest will vary directly as a function of the assessment goals themselves. Additionally, pragmatic considerations (e.g., time and resources) can greatly affect the choice of method employed in the assessment process.

Drawing Inferences

The data derived from the assessment process in experimental psychopathology can be interpreted in distinct ways; the challenge is isolating the best possible information for maximum explanatory value. There are three commonly employed forms of inference in experimental psychopathology research: person-referenced, criterion-referenced, and norm-referenced approaches (Kazdin, 1977).

Person-referenced approaches focus on the individual and compare measured responses to the same person (e.g., number of times of marijuana use per week). The referent is the person himself or herself and his or her own behavior in a particular epoch. Criterion-referenced approaches focus on responses of the individual in the context of a specified standard (e.g., endorsing a score of 10 or higher on a designated alcohol use measure is suggestive of alcohol abuse or dependence). Although criterion-referenced approaches often provide a specific benchmark upon which to evaluate a response, the challenge for most cases of psychopathology has often been in isolating objective indices of "adaptive" responding. Norm-referenced approaches compare the observed responses to a normative group. For example, a researcher may compare the degree and type of attentional bias for threat experienced by a person with generalized anxiety disorder, who is also currently depressed, to the typical attentional biases observed among nondepressed persons with this same anxiety disorder diagnosis.

Determining Assessment Value

With the consideration of the types of inference modalities described above, it is important to note that the quality of the data derived from any given assessment activity of experimental psychopathology research can be interpreted from distinct conceptual models. Just as the goals of the assessment often affect the types of content and methods used, the modes of evaluating the quality of data derived from any given assessment activity vary greatly. These approaches differ in the assumptions made about underlying psychopathology, measurement processes, and interpretation guidelines. Thus, the utilization of any given model for any given instance of experimental psychopathology research may depend on any number of factors (e.g., familiarity with a particular model; agreement and understanding of underlying assumptions).

Arguably the model most commonly employed in experimental psychopathology research is the "classic" psychometric model (Guion, 1980). The basic premise of the psychometric model is that there is measurement error; the goal, therefore, is to develop and utilize instruments that maximize accuracy and minimize error. This approach emphasizes the validity and reliability of a particular assessment tool in capturing the processes or variables of interest. The psychometric model has driven many of the assessment approaches used in better understanding psychopathology. The generalizability

model focuses on determining the nature of variability in regard to the setting or context in which it was obtained (Cone, 1978). In short, variability is understood in relation to the contextual conditions (e.g., time of assessment, setting). To the extent there are large differences in context for any given assessment (e.g., responding to drug cues when in a positive vs. negative mood state), interpretation of those data is made in concert with the context in which it was obtained. The accuracy model posits that the usefulness of a given assessment tool centers on how well it captures the process in question (Cone, 1978). Although seemingly simple, it is often not a pragmatically feasible approach for experimental psychopathology research, as there are so many instances wherein there exists a "standard" to which evaluate "accuracy."

Summary

Each type of experimental psychopathology research involves a consideration of a number of overarching conceptual considerations from an assessment standpoint. There is no single formula or set of standards that will uniformly be applied to all sets of questions being addressed. The assessment approach taken in experimental psychopathology, therefore, is theoretically driven and tied directly to the psychopathology process in question.

Key Challenges to Experimental Psychopathology Research

Although experimental psychopathology offers a unique and powerful lens through which to examine psychopathological processes (Zvolensky et al., 2001), there are numerous challenges—some theoretical and some practical—to the application and overall developmental sophistication of experimental psychopathology research. For example, there are difficulties inherent in the types of laboratory models that can be utilized (Abramson & Seligman, 1977). Specifically, there are notable challenges such as ethics and knowledge regarding (a) the relative ability to comprehensively understand the types of symptoms that characterize a phenotype of interest and (b) the ability to take the steps to produce these symptoms in humans when they are determined (Abramson & Seligman, 1977). Isolating ways to address these types of limitations will be beneficial, and perhaps central, in maximizing the impact of experimental psychopathology on the field as a whole. We now present some of these challenges and, where possible, potential strategies for overcoming them.

Interconnection with Practice

Scholars have frequently lamented the gaps between science and practice (e.g., Onken & Bootzin, 1998). Indeed, the field of behavioral science, as a whole, has made many efforts to call attention to the benefits of such endeavors and devised strategies for doing so (e.g., National Advisory Mental Health Council Behavioral Science Workgroup, 2000). The lack of integration of experimental psychopathology research into mainstream clinical science and practice is highly similar to the widely noted gap between clinical science and clinical practice. Although several factors are likely responsible for this gap, there are two that appear to be at the crux of the matter. We have already alluded to the different analytic agendas of the basic experimental psychopathology researcher and the applied practitioner. This difference is compounded by a second issue, namely a language gap. The scientist prefers a language that is precise, but not necessarily broad in scope, whereas the practitioner prefers concepts that are often broad, but technically imprecise. For instance, emotions are frequently discussed clinically and are often the focus of therapeutic attention, yet emotion and related terms used to describe feelings are not considered technical concepts within the science of psychology. As others have identified, information between diverse fields of psychology needs to be bridged to maximize the full range of possible growth and practical impact of psychological science (Onken & Bootzin, 1998).

Unfortunately, there has not been a systematic *a priori* research agenda focused on understanding how basic biobehavioral processes are altered in specific forms of psychopathology and the implications of these alterations for etiology, course, and successful treatment/prevention. In our view, a sophisticated "translational focus" will require that basic research on biobehavioral processes directly inform clinical questions, and conversely, that observations about naturally occurring psychopathology be used to guide efforts to understand operative biobehavioral processes that give rise to diagnostic labels or more generally "psychopathology." This kind of translational integration will not likely come about via current common practices of providing only parenthetical references to basic research in clinical articles, and similarly when clinical issues are discussed only tangentially in basic research articles. Such efforts do not embody the spirit of experimental psychopathology, for which the core strengths derive from an *a priori* focus on basic research on psychopathology processes with a least one eye on practical utility.

Serious concerns regarding experimental psychopathology research will rightly continue so long as the methods and language employed are not considered in terms of understanding salient and manipulable variables and processes that characterize psychopathology and human suffering more generally.

One solution might involve steps consistent with Hayes' (1987) mutual interest model, whereby basic experimental psychopathologists and applied practitioners collaborate and communicate with one another when their interests intersect. Based upon similar logic, another solution might include the creation of translational research centers, where basic and applied psychopathologists are housed under one roof, devise integrative basic and applied programs of research, outline both technical and nontechnical analyses of the relation between basic processes and psychosocial treatment development and testing, and ultimately disseminate such work to practitioners and the general public. We envision that such dissemination efforts would include both technical and nontechnical descriptions of variables and processes shown to cause and exacerbate psychopathology, and similarly descriptions of how such variables and processes relate to intervention components contained in psychosocial treatments. Ideally, such work would link psychosocial interventions with manipulable variables and processes shown to cause or exacerbate psychopathology. It is our belief that such work would likely yield more powerful psychosocial interventions that focus less on psychiatric diagnosis (i.e., symptom topography) and more on shared vulnerabilities and core processes responsible for human suffering and its alleviation. The resulting treatments would, in principle, become more functional and process-driven.

Process-Oriented Experimental Psychopathology Across Response Systems

Contemporary mental health research and practice are predicated upon accurate classification of psychological disorders. In fact, the reliance on the DSM system is apparent from all standpoints of mental health research and practice. Building upon changes in health care policies and procedures, there has been a well-reasoned "push" to standardize and manualize psychosocial and pharmacological treatments for psychological disorders. Whereas the early study of psychopathology dealt with core biobehavioral processes, one could argue this movement embodies a "return to a medical model" type of thinking (Follette, Houts, & Hayes, 1992; Hayes & Follette, 1992).

Whereas many medical diagnoses implicate pathogenic disease *processes* that underlie medical syndromes, psychopathology research has not always elucidated or emphasized these core processes. Rather, diagnoses themselves have come to dominate contemporary mental health research and practice. For example, behavior problems often are defined by symptoms (i.e., topography) without reference to underlying biobehavioral processes (i.e., functions). Thus, a specific treatment for Disorder X may be "fitted" to a specific patient with Disorder X. Although this type of diagnosis-based clinical research and practice is certainly here to stay for political, pragmatic, and clinical reasons, greater attention to core processes will undoubtedly be important to move psychopathology research forward in a meaningful way (Eifert et al., 1998).

Research and practice that does not attend to core biobehavioral processes may lead to a number of problems. For example, as described by Kiesler (1966) and Persons (1991), there can be a "myth of uniformity," such that all persons with Disorder X are *presumed* to be more alike than different. Clinicians largely operate from within an idiographic framework when working with their clients and rightly question work based exclusively on a DSM type of system because of the implicit nomothetic assumption of uniformity across persons. This may be particularly true for behavioral problems that do not fit within the current DSM system (e.g., Eifert, Zvolensky, & Lejuez, 2000). Moreover, as Wolpe (1989) noted, common problems (e.g., panic disorder) do not necessarily imply common histories, let alone common active clinical processes. In fact, there are often different dysfunctional processes that are operating for two persons with the same disorder (i.e., equifinality) and quite possibly similar behavioral processes that cut across dissimilar DSM categories (i.e., multifinality). Thus, any psychosocial treatment that involves targeting core processes linked to certain historical antecedents will optimally need to address the dysfunctional processes at work for a single case. Experimental psychopathology is well suited for clinical application because it often focuses on core processes that give rise to psychopathology. Thus, it can give clinicians a tangible strategy based upon a theoretical understanding of the dysfunctional processes at work for any one client.

Another, related way in which experimental psychopathology can have an impact on the process level is by contributing to future interdisciplinary movements within behavioral science. In current

practice, it is common for psychopathologists to develop theories about behavior problems with little reference to whether observations are supported by theories at more basic levels of science (e.g., neurobiological). Unfortunately, this results in discontinuity between and within various scientific fields. Reference to this lower level of analysis can and should inform and constrain theory about observed psychopathology at higher levels. At the same time, some may suggest that the "Decade of the Brain" has fueled the viewpoint that all psychopathology outcomes can be reduced to biology. Yet psychopathology cannot be usefully considered strictly in biological terms as by definition it is not reducible to biological processes alone. For example, fear is a functional state characterized by collateral changes across systems and therefore is not reducible to biological activities alone. As Miller and Keller (2000, p. 213) have argued, "we advocate not that every study employ both psychological and biological methods, but that researchers not ignore or dismiss relevant literature, particularly in the conceptualization of their research."

Cross-level of analysis theory development and evaluation of core processes requires broad assessment of pertinent constructs that integrates information from multiple response systems at these different levels of analysis. Utilization of the laboratory context for the examination of clinical psychopathology provides the experimental psychopathology researcher with the necessary flexibility in measurement. The multimethod assessment strategy may be particularly helpful when completed within the context of experimental elicitation of important clinical phenomena. Assessment of emotional response provides one such example. Without broad measurement, any one index may yield ambiguous, incomplete, or misleading information about the affective response (Cacioppo & Tassinary, 1990). Moreover, during clinically relevant cognitive-affective distress states, "differential" information from response domains may reliably manifest to inform theory about underlying mechanisms.

Importantly, such efforts to study psychopathology can be greatly aided by technological advancements, as reflected by those in human neuroimaging. Numerous functional brain imaging techniques are currently available to examine neural mechanisms in clinical psychopathology. Some examples include positron emission tomography, single photon emission computed tomography, and functional magnetic resonance imaging. For example, neuroimaging techniques have contributed significantly to the understanding of the pathophysiology in obsessive-compulsive disorder (Saxena et al., 1998). Thus, experimental psychopathology researchers' ability to advance understanding of the mechanisms responsible for psychopathology increases with the continued application of these brain imaging technologies.

Laboratory Preparations Are Not Equivalent to Clinical Processes

Experimental psychopathologists have devised numerous laboratory preparations to induce and elucidate clinically relevant processes. Examples of such work include preparations that use noncontingent aversive stimulation to study motivation and emotion dysregulation (Mineka & Hendersen, 1985); procedures to elucidate the role of onset and offset predictability and controllability in relation to anxious responding (Zvolensky, Eifert, & Lejuez, 2001); mood induction preparations in depressive responding (Martin, 1990); conditioning preparations to study the acquisition and transfer of aversive affective states to a variety of stimuli (Forsyth, Daleiden, & Chorpita, 2000); and preparations that give rise to verbal-regulatory (i.e., cognitive) processes involved in psychopathology (Hayes, Jacobson, Follette, & Dougher, 1994). Such preparations are often a reliable way to establish or evoke "clinical" processes. Yet experimental preparations are not equivalent with clinical processes, nor should they be treated as such. Indeed, the distinction between procedure and process is critical when considering the clinical relevance of findings from experimental psychopathology. Thus, the reporter of such work has to use cautious language when disseminating this type of scholarship.

There has been a great tendency to equate experimental preparations with clinical processes, leading to erroneous conclusions. Here, we consider Pavlovian or respondent conditioning in relation to understanding the origins and maintenance of anxiety-related disorders as one example of this problem. In its most basic form, such preparations involve pairing a previous neutral stimulus (NS) in a contingency with an unpleasant event or aversive unconditioned stimulus (UCS) that is capable of eliciting a strong negative emotional unconditioned response (UCR). With suitable controls, this preparation will result in a change in the emotion-eliciting functions of the NS, such that it becomes a conditional stimulus (CS) capable of eliciting a conditioned emotional response (CER or CR; i.e., fear and anxiety) more or less similar to the original

UCR (Mineka & Zinbarg, 1996). At the core, these developments share one common etiological thread: aversive stimulus functions are transferred via an association between some otherwise nonthreatening object or event and an abrupt and highly aversive neurobiological response (Forsyth & Eifert, 1998).

Associative learning is not predicated on identifying or manipulating a pain- or fear-inducing stimulus (i.e., a UCS) as is typical of laboratory preparations of associative fear conditioning. For example, operant preparations can yield respondent processes (e.g., punishing consequences can establish antecedents as conditioned suppressors that elicit a wide range of conditioned emotional responses). The critical process variable that enables transfer of aversive stimulus functions to otherwise innocuous stimuli is the response, not identifying contiguous NS–UCS pairings; this view has received recent experimental support (see Forsyth et al., 2000).

Unfortunately, laboratory conditioning preparations involving NS–UCS pairings have been taken as the definitive working model to explain associative processes in the etiology of phobias and other fear-related conditions seen clinically. Hence, if a sequence of events leading to clinical fear onset cannot be construed in terms of pairings between some traumatic/painful UCS in relation to some object or event in the environment, the phobia cannot be due to Pavlovian conditioning (e.g., see Menzies & Clarke, 1995; Rachman, 1991). Here, it is not disputed that phobias may be acquired by means other than direct traumatic conditioning, or more importantly exposure to an identifiable pain-producing aversive event. What is disputed, however, are the contentions that (a) finding an identifiable UCS is the only evidence for direct conditioning and (b) laboratory fear conditioning preparations involving CSs and UCSs are the way to define associative fear onset processes clinically.

As Eysenck (1987) correctly pointed out, from the experimenter or clinician's perspective, evidence for direct conditioning typically involves either the manipulation or identification of neutral stimuli (NSs) in relation to identifiable pain-producing stimuli (UCSs). That is, experimenters tend to define conditioning processes in terms of conditioning preparations. Eysenck goes on to say, however, that from the individual's perspective, direct conditioning involves the experience of abrupt and aversive interoceptive or bodily responses. That is, as far as research subjects and clients are concerned, conditioning involves the bodily effects of UCSs (not UCSs themselves) in relation to objects or events in the environment.

Experimental psychopathologists have emphasized clinically relevant processes and have devised powerful experimental preparations to elucidate such processes. However, Pavlovian conditioning itself was never considered a pathogenic process (Pavlov, 1903), but rather could become pathogenic when interacting with other variables. At some point, Pavlovian fear conditioning began to be viewed as pathogenic itself. Researchers then began to treat respondent conditioning preparations and associative processes as monotypic (Lazarus, 1971; Rachman, 1977). This view has arguably had the unfortunate effect of obscuring clinically relevant learning processes that are involved in the acquisition and maintenance of fearful and anxious behavior seen clinically (Davey, 2002).

What accounts for an otherwise adaptive conditioned emotional response leading to anxiety psychopathology in some individuals, but not others? In this example, a narrow focus on the preparations involved has led to the following spurious conclusions: (a) many persons exposed to events that could be construed in terms of Pavlovian fear conditioning preparations (b) fail to develop clinically significant conditioned emotional responses as a result, and therefore (c) conditioning cannot account for fear onset in the majority of cases seen clinically. A focus on the formal and structural properties of the preparations can either potentiate or depotentiate the probability of conditioning and the extent to which conditioning processes become problematic. Indeed, we know that prior history of control over aversive events, prior history of exposure to stimuli without aversive consequences, and contextual factors can, either alone or in combination, influence whether conditioned emotional responses are acquired and the extent to which they are evoked on subsequent occasions (Bouton, 2000). Consideration of such factors does not mean that conditioning processes are not involved, but rather illustrates that conditioning is complex and functionally determined.

Summary

Experimental psychopathology represents a subfield of psychological science aimed at elucidating the processes underlying abnormal behavior. The present chapter provided a synopsis of the historical perspectives and key elements of experimental psychopathology research. Further, the methodological approaches employed in experimental psychopathology were described in relation to conceptual

considerations. Although experimental psychopathology has made major contributions to the field of psychological science, there are numerous points of entry for it to maximize its integrative potential across basic and applied domains (translational function). Future experimental psychopathology work will likely need to continue to develop and expand in innovative ways to overcome key challenges facing it and the field as a whole.

References

Abramson, L. Y., & Seligman, M. E. P. (1977). Modeling psychopathology in the laboratory: History and rationale. In J. P. Maser & M. E. P. Seligman (Eds.), *Psychopathology: Experimental models* (pp. 1–26). San Francisco: W. H. Freeman.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association (1947). Recommended graduate training program in clinical psychology. *American Psychologist*, *2*, 539–558.

Anderson, O. D., & Liddell, H. S. (1935). Observations on experimental neurosis in sheep. *Archives of Neurological Psychiatry*, *34*, 330–354.

Benjamin, L. Jr. (2000). The psychology laboratory at the turn of the 20th century. *American Psychologist*, *55*, 318–321. DOI: 10.1037/0003-066X.55.3.318

Bouton, M. E. (2000). A learning theory perspective on lapse, relapse, and the maintenance of behavior change. *Health Psychology*, *19*, 57–63. DOI: 10.1037/0278-6133.19.Suppl1.57

Brown, J. F. (1937). Psychoanalysis, topological psychology and experimental psychopathology. *Psychoanalytic Quarterly*, *6*, 227–237.

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45*, 16–28. DOI: 10.1037/0003-066X.45.1.16

Chapman, L. J., & Chapman, J. P. (1973). *Disordered thought in schizophrenia*. Englewood Cliffs, NJ: Prentice-Hall.

Clark, L. A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. *Journal of Abnormal Psychology*, *103*, 103–116. DOI: 10.1037/0021-843X.103.1.103

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy*, *9*(5), 882–888. DOI: 10.1016/S0005-7894(78)80020-3

Davey, G. C. L. (2002). "Nonspecific" rather than "nonassociative" pathways to phobias: A commentary on Poulton and Menzies. *Behaviour Research and Therapy*, *40*, 151–158. DOI: 10.1016/S0005-7967(01)00046-8

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*, 4, 662–680. DOI: 10.1037/0021-9010.70.4.662

Eifert, G. H., Schulte, D., Zvolensky, M. J., Lejuez, C. W., & Lau, A. W. (1998). Manualized behavior therapy: Merits and challenges. *Behavior Therapy*, *28*, 499–509. DOI: 0005-7894/97/0499-050951.0

Eifert, G. H., Zvolensky, M. J., & Lejuez, C. W. (2000). Heart-focused anxiety and chest pain: A conceptual and clinical review. *Clinical Psychology: Science and Practice*, *7*(4), 403–417. DOI: 10.1093/clipsy/7.4.403

Estes, W. K., & Skinner, B. F. (1941). Some quantitative properties of anxiety. *Journal of Experimental Psychology*, *29*, 390–400. DOI: 10.1037/h0062283

Eysenck, H. J. (Ed.) (1961). *Handbook of abnormal psychology: An experimental approach*. New York: Basic Books.

Eysenck, H. J. (Ed.) (1973). *Handbook of abnormal psychology*. San Diego, CA: EdITS Publishers.

Eysenck, H. J. (1987). Behavior therapy. In H. J. Eysenck & I. Martin (Eds.), *Theoretical foundations of behavior therapy* (pp. 3–34). New York: Plenum.

Felmingham, K., Williams, L. M., Kemp, A. H., Liddell, B., Falconer, E., Peduto, A., &, Bryant, R. (2010). Neural responses to masked fear faces: Sex differences and trauma exposure in posttraumatic stress disorder. *Journal of Abnormal Psychology*, *119*(1), 241–247. DOI: 10.1037/a0017551.

Follette, W. C., Houts, A. C., & Hayes, S. C. (1992). Behavior therapy and the new medical model. *Behavioral Assessment*, *14*, 323–343.

Forsyth, J. P., Daleiden, E. L., & Chorpita, B. F. (2000). Response primacy in fear conditioning: Disentangling the contributions of UCS vs. UCR intensity. *Psychological Record*, *50*, 17–33.

Forsyth, J. P., & Eifert, G. H. (1998). Response intensity in content-specific fear conditioning comparing 20% versus 13% CO2-enriched air as unconditioned stimuli. *Journal of Abnormal Psychology*, *107*(2), 291–304. DOI: 10.1037/0021-843X.107.2.291

Forsyth, J. P., & Zvolensky, M. J. (2002). Experimental psychopathology, clinical science, and practice: An irrelevant or indispensable alliance? *Applied and Preventive Psychology: Current Scientific Perspectives*, *10*, 243–264. DOI: 10.1016/S0962–1849(01)80002–0

Franks, C. M. (Ed.) (1964). *Conditioning techniques in clinical practice and research*. Berlin: Springer.

Franz, S. I. (1912). Experimental psychopathology. *Psychological Bulletin*, *9*, 145–154.

Gantt, W. H. (1971). Experimental basis for neurotic behavior. In H. D. Kimmel (Ed.), *Experimental psychopathology: Recent research and theory* (pp. 33–48). New York: Academic Press.

Gregor, A. A. (1910). *Leitfaden der experimentellen psychopathologie*. Berlin: Allzeit Voran.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, *11*(3), 385–398. DOI: 10.1037/0735-7028.11.3.385

Hawkins, K. A., & Cougle, J. R. (2011). Anger problems across the anxiety disorders: Findings from a population-based study. *Depression and Anxiety*, *28*, 145–152. DOI: 10.1002/da.20764

Hayes, S. C. (1987). The relation between "applied" and "basic" psychology. *Behavior Analysis*, *22*, 91–100.

Hayes, S. C., & Follette, W. C. (1992). Can functional analysis provide a substitute for syndromal classification? *Behavioral Assessment*, *14*, 345–365.

Hayes, S. C., Jacobson, N. S., Follette, V. M., & Dougher, M. J. (Eds.) (1994). *Acceptance and change: Content and context in psychotherapy*. Reno, NV: Context Press.

Hoch, P. H., & Zubin, J. (Eds.) (1957). *Experimental psychopathology*. New York: Grune & Stratton.

Hunt, J. M., & Cofer, C. N. (1944). Psychological deficit. In J. M. Hunt (Ed.), *Personality and the behavior disorders* (pp. 971–1032). Oxford: Ronald Press.

Ingram, R. E. (1986). *Information processing approaches to clinical psychology*. Orlando, FL: Academic Press.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*(1), 141–150.

Kazdin, A. E. (1982). Single-case experimental designs in clinical research and practice. *New Directions for Methodology of Social & Behavioral Science*, *13*, 33–47.

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, *65*, 110–136. DOI: 10.1037/h0022911

Kihlstrom, J. F., & McGlynn, S. M. (1991). Experimental research in clinical psychology. In M. Hersen, A. Kazdin, & A. Bellack (Eds.), *Clinical psychology handbook* (pp. 239–257). New York: Pergamon Press.

Kimmel, H. D. (1971). Introduction. In H. D. Kimmel (Ed.), *Experimental psychopathology: Recent research and theory* (pp. 1–10). New York: Academic Press.

Krasnogorski, N. I. (1925). The conditioned reflexes and children's neuroses. *American Journal of Disorders of Children*, *30*, 753–768.

Landis, C. (1949). Experimental methods in psychopathology. *Mental Hygiene*, *33*, 96–107.

Lazarus, A. A. (1971). *Behavior therapy and beyond*. New York: McGraw-Hill.

Lenzenweger, M. F., & Dworkin, R. H. (Eds.) (1998). *Origins and development of schizophrenia: Advances in experimental psychopathology*. Washington, DC: American Psychological Association.

Liddell, H. S. (1938). The experimental neurosis and the problem of mental disorder. *American Journal of Psychiatry*, *94*, 1035–1041.

Lilienfeld, S. O. (1996, Jan/Feb). EMDR treatment: Less than meets the eye? *Skeptical Inquirer*, 25–31.

Lubin, A. J. (1943). The experimental neurosis in animal and man. *American Journal of the Medical Sciences*, *205*, 269–277. DOI: 10.1097/00000441-194302000-00026

Mackinnon, D. W., & Henle, M. (1948). *Experimental studies in psychodynamics; A laboratory manual*. Cambridge, MA: Harvard University Press.

Martin, M. (1990). On the induction of mood. *Clinical Psychology Review*, *10*, 669–697. DOI: 10.1016/0272-7358(90)90075-L

Maser, J. D., & Seligman, M. E. P. (1977). *Psychopathology: Experimental models*. San Francisco: W. H. Freeman.

Masserman, J. H. (1943). Experimental neuroses and psychotherapy. *Archives of Neurology and Psychiatry*, *49*, 43–48.

McFall, R. M. (1991). Manifesto for a science of clinical psychology. *Clinical Psychologist*, *44*(6), 75–88.

McNally, R. J. (1998). Information-processing abnormalities in anxiety disorders: Implications for cognitive neuroscience. *Cognition and Emotion*, *12*, 479–495. 10.1080/026999398379682

Menzies, R. G., & Clarke, J. C. (1995). The etiology of phobias: A non-associative account. *Clinical Psychology Review*, *15*, 23–48. 10.1016/0272-7358(94)00039-5

Miller, G., & Keller, J. (2000). Psychology and neuroscience: Making peace. *Current Directions in Psychological Science*, *9*, 212–215. DOI: 10.1111/1467-8721.00097

Mineka, S., & Hendersen, R. W. (1985). Controllability and predictability in acquired motivation. *Annual Review of Psychology*, *36*, 495–529. DOI: 10.1146/annurev.ps.36.020185.002431

Mineka, S., & Zinbarg, R. (1996). Conditioning and ethological models of anxiety disorders: Stress-in-dynamic context anxiety models. In D. A. Hope (Ed.), *Perspectives on anxiety, panic, and fear: Volume 43 of the Nebraska Symposium on Motivation* (pp. 135–210). Lincoln, NB: Nebraska University Press.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387. DOI: 10.1037/0003-066X.38.4.379

National Advisory Mental Health Council Behavioral Science Workgroup (2000). *Translating behavioral science into action*. Washington, DC: National Institutes of Health.

Onken, L. S., & Bootzin, R. R. (1998). Behavioral therapy development and psychological science: If a tree falls in the forest and no one hears it. *Behavior Therapy*, *29*, 539–544.

Osgood, C. E. (1953). *Method and theory in experimental psychology*. New York: Oxford University Press.

Pavlov, I. P. (1903). Experimental psychology and psychopathology in animals. *Herald of the Military Medical Academy*, *7*(2), 109–121.

Pavlov, I. P. (1961). *Psychopathology and psychiatry: I. P. Pavlov selected works*. San Francisco: Foreign Languages Publishing House.

Persons, J. B. (1991). Psychotherapy outcome studies do not accurately represent current models of psychopathology. *American Psychologist*, *46*, 99–106. DOI: 10.1037/0003-066X.46.2.99

Peters, K. D., Constans, J. I., & Mathews, A. (2011). Experimental modification of attribution processes. *Journal of Abnormal Psychology*, *120*(1), 168–173. DOI: 10.1037/a0021899

Popplestone, J. A., & McPherson, M. W. (1984). Pioneer psychology laboratories in clinical settings. In J. Brozek (Ed.), *Explorations in the history of psychology in the United States* (pp. 196–272). Lewisburg, PA: Bucknell University Press.

Proctor, C., Maltby, J., & Linley, A. P. (2011). Strengths use as a predictor of well-being and health-related quality of life. *Journal of Happiness Studies*, *12*, 1, 153–169. DOI: 10.1007/s10902-009-9181-2

Rachman, S. (1966). Sexual fetishism: An experimental analogue. *Psychological Record*, *16*, 293–296.

Rachman, S. (1977). The conditioning theory of fear acquisition: A critical examination. *Behaviour Research and Therapy*, *15*, 375–387. DOI: 10.1016/0005-7967(77)90041-9

Rachman, S. (1991). Neo-conditioning and the classical theory of fear acquisition. *Clinical Psychology Review*, *11*, 155–173. DOI: 10.1016/0272-7358(91)90093-A

Sandler, J., & Davidson, R. S. (1971). Psychopathology: An analysis of response consequences. In H. D. Kimmel (Ed.), *Experimental psychopathology: Recent research and theory* (pp. 71–93). New York: Academic Press.

Saxena, S., Brody, A., Schwartz, J., & Baxter, L. (1998). Neuroimaging and frontal-subcortical circuitry in obsessive-compulsive disorder. *British Journal of Psychiatry*, *173*, 26–37.

Schumann, F. (1905). Proceedings of the First Congress of Experimental Psychology, at Giessen, April, 1904. *Psychological Bulletin*, *2*, 81–86.

Shenger-Krestovnikova, N. R. (1921). Contributions to the question of differentiation of visual stimuli and the limits of differentiation by the visual analyzer of the dog. *Bulletin of the Lesgaft Institute of Petrograd*, *3*, 1–43.

Sher, K. J., & Trull, T. J. (1996). Methodological issues in psychopathology research. *Annual Review of Psychology*, *47*, 371–400. DOI: 10.1146/annurev.psych.47.1.371

Skinner, B. F. (1953). *Science and human behavior*. New York: The Free Press.

Taylor, E. (1996). *William James on consciousness beyond the margin*. Princeton, NJ: Princeton University Press.

Wells, F. L. (1914). Experimental psychopathology. *Psychological Bulletin*, *11*, 202–212. DOI: 10.1037/h0073486

Williams, J. M. G., Mathews, A., & MacLeod, C. (1997). The emotional Stroop task and psychopathology. *Psychological Bulletin*, *120*, 3–24.

Wolpe, J. (1952). Experimental neuroses as learned behavior. *British Journal of Psychology*, *43*, 243–268.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA: Stanford University Press.

Wolpe, J. (1989). The derailment of behavior therapy: A tale of conceptual misdirection. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*, 3–15. DOI: 10.1016/0005-7916(89)90003-7

Wolpe, J., Salter, A., & Reyna, L. J. (Eds.) (1964). *The conditioning therapies*. New York: Holt, Rinehart.

Yerofeeva, M. N. (1912). *Electrical stimulation of the skin of the dog as a conditioned salivary stimulus.* Unpublished thesis.

Yerofeeva, M. N. (1916). Contribution to the study of destructive conditioned reflexes. *Comptes Rendus de la Societé Biologique*, *79*, 239–240.

Zubin, J., & Hunt, H. F. (1967). *Comparative psychopathology, animal and human*. New York: Grune and Stratton.

Zvolensky, M. J., & Eifert, G. H. (2000). A review of psychological factors/processes affecting anxious responding during voluntary hyperventilation and inhalations of carbon dioxide-enriched air. *Clinical Psychology Review*, *21*, 375–400. DOI: 10.1016/S0272-7358(99)00053-7

Zvolensky, M. J., Eifert, G. H., & Lejuez, C. W. (2001). Emotional control during recurrent 20% carbon dioxide-enriched air induction: Relation to individual difference variables. *Emotion*, *2*, 148–165. DOI: 10.1037//1528-3542.1.2.148

Zvolensky, M. J., Lejuez, C. W., Stuart, G. L., & Curtin, J. J. (2001). Experimental psychopathology in psychological science. *Review of General Psychology*, *5*, 371–381. DOI: 10.1037/1089-2680.5.4.371

CHAPTER

3 Single-Case Experimental Designs and Small Pilot Trial Designs

Kaitlin P. Gallo, Jonathan S. Comer, *and* David H. Barlow

Abstract

This chapter covers single-case experimental designs and small pilot trial designs, beginning with a review of the history of single-case experimental designs. Such designs can play key roles in each stage of treatment development and evaluation. During the earliest stages of treatment development and testing, single-case experimental designs clarify functional relationships between treatment and symptoms. After a treatment has been formalized, a series of replicating single-case experiments in conjunction with randomized clinical trials can contribute meaningful information to efficacy evaluations. After a treatment has demonstrated robust efficacy in large-scale clinical trials, single-case designs can speak to the generalizability and transportability of treatment efficacy by demonstrating the successful application of established treatments when flexibly applied to individuals, or in settings, that may vary in important and meaningful ways. Specific designs covered include A-B designs, basic withdrawal designs (i.e., A-B-A trials), extensions of the traditional withdrawal design (e.g., A-B-A-B designs, B-A-B-A designs, and A-B-C-B designs), multiple-baseline trials, and small pilot trial designs, all of which assess treatment effects in a systematic manner with a relatively small number of participants. We conclude with a call for increased utilization of single-case experimental designs in clinical psychology treatment outcomes research.

Key Words: Single-case experimental designs, multiple-baseline designs, withdrawal designs, treatment outcomes research, idiographic and nomothetic group design evaluations

Evidence-based practice in clinical psychology entails an explicit and judicious integration of best available research with clinical expertise, in the context of client characteristics, preferences, and values. Such an endeavor necessitates a compelling body of evidence from which to draw. Systematic, carefully designed treatment evaluations—the cornerstone of applied psychology—are central to this pursuit and allow data to meaningfully influence individual clinical practice, mental health care debates, and public policy. Large controlled group comparison designs as well as experimental designs utilizing only a few participants each contribute valuable evidence in this regard. In this chapter, we address the latter set of designs—those designs that can provide systematic and rich evidence of treatment effects with a relatively small number of participants.

The National Institute of Drug Abuse commissioned a report broadly outlining a sequence of three progressive stages of treatment development and evaluation (Barlow, Nock, & Hersen, 2009; Kazdin, 2001). In *Stage 1*, the first phase of treatment development and testing, novel interventions eventuate from a scholarly integration of theory, previous research, and consultation with relevant experts (the formal process of treatment development is covered elsewhere; Rounsaville, Carroll, & Onken, 2001). To provide preliminary evidence that the intervention is associated with meaningful change, pilot testing is conducted on a relatively

small number of participants who are representative of the population of clients for whom the treatment is designed. Stage 1 research activities afford opportunities to refine treatment procedures as necessary prior to large-scale treatment evaluation in response to early data, and to focus on key preliminary issues related to treatment feasibility, tolerability, and credibility and consumer satisfaction.

Once an intervention has been formalized and feasibility and preliminary efficacy have been established, *Stage 2* research entails larger-scale evaluations—typically group comparisons in tightly controlled trials—to firmly establish treatment efficacy and to evaluate potential mediators and moderators of treatment response. *Stage 3* consists of research efforts to evaluate the broad effectiveness and transportability of outcomes demonstrated in Stage 2 laboratory studies to less controlled practice settings.

From a methodology and design perspective, Stage 1 research is typically the purview of idiographic single-case experimental designs and small pilot randomized controlled trials (RCTs), whereas Stage 2 activities are addressed with adequately powered RCTs, nomothetic group comparisons, and formal tests of mediation and moderation (see Kendall, Comer, & Chow, this volume; see also MacKinnon, Lockhart, & Gelfand, this volume). Stage 3 research activities utilize a diversity of designs, including single-case designs, RCTs, sequential multiple assignment randomized trial (SMART) designs (Landsverk, Brown, Rolls Reutz, Palinkas, & Horwitz, 2011), practical clinical trials (March et al., 2005), qualitative methods, and clinical epidemiology to address the transportability of treatment effects and the uptake of supported practices in community settings (see Beidas et al., this volume).

In this chapter we cover single-case experimental designs, including A-B designs, basic withdrawal designs (i.e., A-B-A trials), extensions of the traditional withdrawal design (e.g., A-B-A-B designs, B-A-B-A designs, and A-B-C-B designs), and multiple-baseline trials. As Barlow, Nock, and Hersen (2009) note, single-case experimental designs can play key roles in each of the three outlined stages of treatment evaluation. Such designs are essential for initial Stage 1 evaluations, clarifying functional relationships between treatment and symptoms. After a treatment has been formalized, a series of replicating single-case experiments can contribute meaningfully to Stage 2 efficacy evaluations. After a treatment has demonstrated robust efficacy in large-scale RCTs, single-case designs contribute to Stage 3 efforts by demonstrating the successful application of established treatments when flexibly applied to individuals, or in settings, that may vary in important and meaningful ways (e.g., Suveg, Comer, Furr, & Kendall, 2006). Accordingly, in many ways comprehensive treatment evaluation begins and ends with the study of change in the individual. In this chapter we also cover small pilot trial designs, which formally set the stage for Stage 2 research. Collectively, the designs covered in this chapter all share the advantage of providing systematic and compelling evidence of treatment effects with a relatively small number of participants.

We begin with a brief historical overview of the role of single-case designs in clinical psychology, followed by consideration of some general procedures, and then examine the prototypical single-case experimental designs including multiple-baseline designs. We then examine key methodological and design issues related to the small pilot RCT, which can serve as a bridge from idiographic single-case designs to nomothetic group comparison research, and conclude with a renewed sense of urgency for research utilizing the experimental designs considered in this chapter.

Single-Case Designs: A Brief Historical Overview

Until relatively recently, the field of clinical psychology lacked an adequate methodology for studying individual behavior change. Hersen and Barlow (1976) outlined procedures for studying changes in individual behavior, with foundations in laboratory methods in experimental physiology and psychology. Prior to this emergence of systematic procedures, less robust procedures dominated the field of applied clinical research, including the popular but less scientific case study method (Bolger, 1965) that dominated clinical psychology research for the first half of the twentieth century. These case studies tended to be relatively uncontrolled and researchers often drew expansive conclusions from their data, with some exceptions (e.g., Watson & Rayner, 1920).

In the middle of the twentieth century, an increased focus on more rigorously applied research and statistical methods fueled a split between those investigators who remained loyal to uncontrolled case studies (which, despite frequent exaggerated conclusions of a treatment's efficacy, often contained useful information about individual behaviors) versus investigators who favored research that compared differences between groups. By the

late 1940s, some clinical researchers started using between-subjects group designs with operationalized dependent variables (Barlow et al., 2009). Although these early efforts (e.g., Barron & Leary, 1955; Powers & Witmer, 1951) were crude by today's standards and the most usual result was "no differences" between therapy and comparison group, the idea that therapeutic efficacy must be established scientifically slowly took hold. This notion was reinforced by Eysenck's (1952) controversial conclusion (based on limited studies and actuarial tables) that untreated patients tended to improve as much as those assumed to be receiving psychotherapy.

Despite the increase in the popularity of between-group comparisons in the latter part of the twentieth century, several factors impeded its utility and impact for the first few decades of its use (Barlow et al., 2009). For example, some clinicians worried that withholding treatment for those study participants assigned to a comparison group might be unethical. Practically, researchers found it difficult to recruit sufficient populations of people with low-base rate disorders for their studies (an issue that has improved with the advent of the multisite clinical trial). Results were typically presented in an averaged or aggregated format, obscuring within-subject variability and decreasing the generalizability of the findings.

Clinical investigators have begun to debate the merits of idiographic and nomothetic approaches to treatment evaluation (Barlow & Nock, 2009). Evaluation of dependent variables comparing averaged data from large groups of people (nomothetic approach) is an essential method with which to establish treatment efficacy and effectiveness, and with which to inform broad public policy. However, the generalizability of data obtained by these approaches may be limited in some cases, as the true effects of the independent variable for individual subjects may be blurred among reported averages. Research designs that examine individuals on a more intensive level (idiographic approach) allow for a more specific understanding of the mechanisms of an intervention and its effects on different presentations as they pertain to the individual, although such methods confer less generalizability relative to nomothetic approaches. The importance of single-case designs is prominently featured in the establishment of consensus clinical guidelines and best practice treatment algorithms (e.g., American Psychological Association, 2002). Following years of debate, we believe that in the modern evidence-based practice landscape both methodological traditions utilizing systematic and tightly controlled designs should play prominent, complementary roles.

General Procedures

In single-case experimental research, a repeated measures design, in which data are collected systematically throughout the baseline and treatment phases, is essential in order to comprehensively evaluate treatment-related change. Although two-point, pre–post measurement strategies can examine the broad impact of an intervention, systematic repeated measurements across an intervention phase allow for a nuanced examination of how, why, and when changes happen (Barlow et al., 2009). Measurements must be specific, observable, and replicable (Kazdin, 2001; Nock & Kurtz, 2005) and are ideally obtained under the same conditions for each observation, with the measurement device and all environmental conditions remaining constant. Specificity of observations refers to measurement precision and the extent to which the boundaries of a target behavior are made clear. For example, a target behavior that calls for a child to demonstrate "appropriate classroom behavior" is less specific than one that calls for a child to "remain seated and not talk out of turn for a period of 60 minutes."

Repeated assessments are critical, but the researcher must carefully balance the need for sufficient information with the need to avoid subject fatigue when determining the frequency of measurements. The researcher must also carefully consider whether to rely only on self-report measures, which can be influenced by social desirability (i.e., the inclination for participants to behave in a way that they think will be perceived well by the experimenter) (Crowne & Marlowe, 1960) and/or demand characteristics (i.e., the change in behavior that can occur when a research participant formulates beliefs about the purpose of the research) (Orne, 1962), or whether to include structured behavioral observations as well. Our view is that clinical researchers should always make attempts to incorporate behavioral observations into single-case experimental designs. Finally, given the small sample size associated with single-case experimental designs, the researcher must take care when interpreting data, especially in the case of extreme variability, so that outliers do not unnecessarily skew results and conclusions (Barlow et al., 2009). This is particularly challenging in the case of nonlinear changes in target behaviors. Experimental phases should be long enough to differentially identify random outliers from systematic cyclic variations

in target outcomes. When nonlinear variations can present challenges to interpretation, the clinical researcher is wise to extend measurement procedures to evaluate whether a steady and stable pattern emerges.

Procedurally, most single-case experimental designs begin with a *baseline* period, in which target behaviors are observed repeatedly for a period of time. The baseline phase, often called the "A" phase, demonstrates the stability of the target behavior prior to the intervention so that the effects of the intervention can be evaluated against the naturalistic occurrence of the target behavior (Risley & Wolf, 1972). Baseline (phase A) observations also provide data that predict future levels of the target behavior. Demonstrating decreasing symptoms after the initiation of treatment may be less compelling if symptoms already were shown to be systematically declining across the baseline period.

Although a stable baseline pattern is preferable, with no variability or slope in the target behavior(s) (Kazdin, 1982, 2003), this may be difficult in applied clinical research (Sidman, 1960). Accordingly, visual inspection and statistical techniques can be utilized to compare phases to each other (Barlow et al., 2009). For example, interrupted time-series analyses (ITSA) allow the researcher to evaluate changes in the slope and level of symptom patterns induced by treatment by first calculating omnibus tests (F statistic) of slope and level changes, with follow-up post hoc t tests to examine which specific aspect was affected by treatment initiation (slope, level, or both). The double bootstrap method (McKnight, McKean, & Huitema, 2000) entails iterative statistical resampling methods to achieve less biased estimates that are particularly well suited for small n single-case experiments.

When moving between phases in single-case experimental research, it is crucial to change only one variable at a time (Barlow et al., 2009). Otherwise, it is impossible to determine which manipulation was responsible for changes in a target behavior. Data stability on a target behavior is widely regarded as a necessary criterion that must be achieved prior to progressing to the next phase (Barlow et al., 2009).

Single-Case Experimental Designs

Having provided a general overview of the major considerations and procedures involved in single-case design research, we now turn our attention to the prototypical single-case experimental designs. We begin with an overview of the major types of single-case experimental designs, with the goal of familiarizing the reader with the merits and limitations of each and providing brief illustrations from published research. Specifically, we consider A-B designs (a bridge between the case study and experimental design) and then move on to the basic withdrawal designs: A-B-A designs, A-B-A-B designs, B-A-B-A designs, and A-B-C-B designs. We follow with a consideration of multiple-baseline designs. Whereas withdrawal designs are marked by the removal of an intervention after behavior change is accomplished, multiple-baseline designs are marked by different lengths of an initial baseline phase, followed by phase changes across people, time, or behaviors.

A-B Designs

Whereas case studies afford opportunities to study infrequently occurring disorders, to illustrate clinical techniques, and to inspire larger systematic clinical trials, such efforts do not afford causal conclusions. Additionally, it is difficult to remove clinical bias from the reported results. Even with repeated assessment (e.g., Nock, Goldman, Wang, & Albano, 2004), internal validity cannot be guaranteed. A-B designs use repeated measurements and as such represent a transition between case studies and experiments (where the independent variable is manipulated), allowing the researcher to systematically examine a variable of interest during an intervention phase of treatment against its value during a baseline period.

In the *traditional A-B design*, a target behavior is identified and then measured repeatedly in the A (baseline) and B (intervention) phases (Hayes, Barlow, & Nelson-Gray, 1999). In the baseline phase, data about the natural (pre-intervention) occurrence of the target behavior are collected. The researcher then introduces the intervention, continues collecting repeated measures, and examines changes in the target behavior.

The *A-B with follow-up* design includes the same components as the A-B design, with the addition of a period of repeated measurements following the B intervention phase. This design provides more evidence of the stability of an intervention's effects than the traditional A-B design; however, it is still possible that improvements seen in the follow-up phase are not the result of the intervention but of some other factor. In instances where multiple behaviors or multiple measures are of interest, an *A-B design with multiple target measures and follow-up* can be utilized. For example, a researcher might collect measures of both anxiety *and* depression across an

A-B design, or might collect measures of a single behavior (such as alcohol use) across multiple settings. A-B designs can also include *a follow-up period and booster treatment* if it becomes clinically indicated during the follow-up period for the B phase, or the intervention, to be briefly reinstated. This is similar to the A-B-A-B design, which we discuss later in this section.

Cooper, Todd, Turner, and Wells (2007) used an A-B design with multiple target measures and follow-up in their examination of cognitive-behavioral treatment for bulimia nervosa. Baseline measurements were relatively stable, with decreases in symptomatology (bingeing, vomiting, and negative beliefs about eating) beginning during the treatment (B) phase and maintained at 3- and 6-month follow-up points. Results were similar for the other two participants. Figure 3.1 shows an example of an A-B design.

The A-B design allows the researcher to avoid some of the drawbacks of the case study approach when examining only one individual. While certainly not the most rigorous of the single-case strategies, the A-B design can be helpful "transitory strategy" in cases where true experimental methods, such as an RCT or a repetition of the A-B phases, are not possible (Campbell & Stanley, 1966; Cook & Campbell, 1979). The major strength of the A-B design is that when the target behavior demonstrates stability during the baseline period and the behavior changes upon intervention, one can infer that the change may have been a result of the intervention (Barlow et al., 2009). However, conclusions from this design are vulnerable to multiple possible threats to internal and external validity; thus, the transitory strategy of the A-B design should be used only when other more systematic methods are not possible (Campbell, 1969).

Despite its clinical utility and improvements over the traditional case study, several limitations hinder the methodological vigor of the A-B design. The biggest problem with this design is that observed changes in the B phase may not be caused by the intervention but instead by some other factor (Wolf & Risley, 1971). As such, Campbell and Stanley (1966) advocate the use of the term "quasi-experimental design" to describe that correlative factors may be just as likely to account for observed change as the intervention itself. In the bulimia treatment study by Cooper and colleagues (2007), although it is certainly possible that treatment caused the improvements seen, it is impossible to confirm this hypothesis given that the design does not control for the possibility that some other variable was responsible for improvements. Additionally, withdrawal designs are meaningful only to the extent that the intervention can be withdrawn (e.g., one can withdraw reinforcement or a drug, but not surgery or a cognitive problem-solving skill).

A-B-A Design

The A-B-A design offers a more rigorous research design than the A-B design. With the A-B-A strategy,

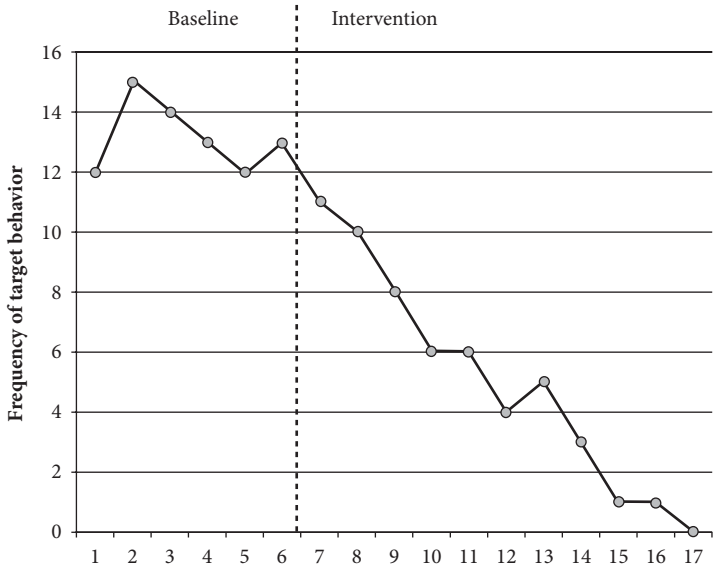


Figure 3.1 Example of an A-B design.

repeated measurements are collected during a baseline period of measurement (A), which is followed by the intervention (B), followed by the withdrawal of the intervention (A). The A-B-A design allows for firmer conclusions than does the A-B design because the effects of the intervention (B phase) can be compared against the effects of the removal of that intervention (in the second phase A, an effective return to baseline). Here, the researcher systematically controls both the introduction and the removal of an independent variable, in this case an intervention. Manipulating an independent variable is the hallmark of a true experiment. If the baseline phase is stable, improvements are then observed in the B phase followed by a return to baseline levels in the second A phase, the experimenter can conclude that the changes likely occurred as a result of the intervention. The certainty with which one can make such a conclusion increases with each replication in different subjects.

Moore and colleagues (Moore, Gilles, McComas, & Symons, 2010) used an A-B-A withdrawal design to evaluate the effects of functional communication training (FCT) on nonsuicidal self-injurious behavior in a male toddler with a traumatic brain injury. FCT involves teaching effective communication strategies that are meant to replace self-injurious or other undesirable behaviors (Moore et al., 2010). The boy in this examination was taught to use a button to communicate with his mother in order to tell her that he would like her to come in the room. In the first (A) phase, when the boy pressed the button, his mother was to give him 10 seconds of positive attention, and when he hurt himself, he was to receive no attention. In the B phase, the opposite contingencies occurred: attention for self-injury but none for the newly learned form of communication. Following the B phase, the A phase was repeated. In this intervention, the toddler's functional communication was markedly higher, and self-injurious behavior markedly lower, in both training (A) phases as compared to the phase when the contingency was removed (B phase). Given the clear shift from A to B and then from B to A, one can conclude with some certainty that the intervention was responsible for the improvements in this case. Figure 3.2 shows an example of an A-B-A design.

Despite the advantages of the A-B-A design over traditional A-B designs, this design concludes on the nontreatment phase, perhaps limiting the full clinical benefit that the subject can receive from the treatment (Barlow & Hersen, 1973). In addition, the A-B-A design has a sequential

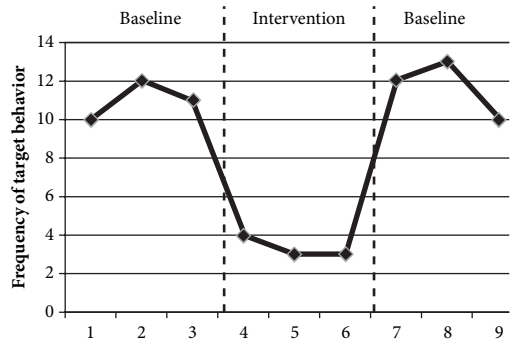


Figure 3.2 Example of an A-B-A design.

confound—specifically, the ordering of treatment introduction may affect the strength of response during the final phase A (Bandura, 1969; Cook & Campbell, 1979; Kazdin, 2003). Additionally, researchers should keep in mind that many withdrawal designs within clinical psychology may be of limited utility, given that "unlearning" clinical skills during the withdrawal phase may be difficult or impossible.

A-B-A-B Design

The A-B-A-B design is often the preferred strategy of the single-case researcher, given its rigorous design and clinical utility. In an A-B-A-B design, repeated measurements are collected through each of four phases: the baseline, the first intervention phase, the second baseline, and then an additional intervention phase. In most published A-B-A-B studies, only one behavior is targeted. However, in some cases, other behaviors that are not a target of the intervention can be measured so that the experimenter can monitor side effects (Kazdin, 1973). Note that this design differs from the A-B design with booster treatment in that the second round of treatment is identical to the first, rather than a simple scaled-down version of the B phase. The A-B-A-B design improves upon the A-B and A-B-A designs by affording increased opportunities to systematically evaluate the link between intervention and target behavior, providing more support for causal conclusions. Essentially, the design affords a built-in replication of findings observed in the initial two study phases. Additionally, from an ethical standpoint, some may prefer to end the experiment on the intervention phase rather than on the withdrawal phase, so that the subject can continue to experience the greatest possible treatment benefits—which may not occur if the individual ends his or her participation in a nonintervention phase.

Importantly, the experimenter cannot control every situation during data collection within an A-B-A-B design. For example, in some clinical circumstances, the phase change may occur at the request (or at the whim) of the subject (e.g., Wallenstein & Nock, 2007). Such an occurrence considerably limits the strength of conclusions due to potential confounding variables that may have led to both changes in the target behavior and the decision to change phases. However, if the A-B-A-B design is followed to fruition, and when study phase changes are controlled entirely by the researcher and not extraneous factors, one can maintain some confidence in the treatment effects. An additional limitation is that if improvements occur during a baseline period, conclusions about the efficacy of the intervention are significantly limited. In such a case, it would behoove the clinician to replicate the examination, either with the same person or additional people who have the same presenting problem.

One of the main limitations of the A-B-A-B design and other withdrawal designs is the experimenter's knowledge of all phase changes and results, which may bias when study phases are changed and how behaviors are evaluated (Barlow et al., 2009). For example, if an experimenter hypothesizes that an intervention will reduce depression, she may change to the intervention phase if the depression starts to remit during the withdrawal phase, or she may wish to keep the intervention phase for a longer period of time than planned if results are not immediately observed. Determining phase lengths in advance eliminates this potential for bias. However, considering clinical response when determining when to switch phases may be important for some research questions, such as when phase changes are to be made after a symptom improvement, or after data have stabilized. For such cases, we recommend that the research develop clear clinical response criteria for phase change prior to initiating the study, and strictly adhere to those criteria to determine phase shifts.

Hunter, Ram, and Ryback (2008) attempted to use an A-B-A-B design to examine the effects of satiation therapy (Marshall, 1979) to curb a 19-year-old man's sexual interest in prepubescent boys—work that illustrates how factors outside of an experimenter's control can in practice affect the intended design. The goal of satiation therapy is to prescribe prolonged masturbation to specific paraphilic fantasies, which is meant to cause boredom and/or extinction of deviant sexual arousal to those paraphilic cues. Phase A in this study consisted of baseline measurement collection, and phase B consisted of satiation therapy. The initial plan called for the schedule to include 14 days of baseline, then 14 days of treatment, followed by an additional 14 days of baseline, and 14 additional days of treatment, with three daily recordings of the dependent variables. In this particular study, 8 days of treatment nonadherence occurred at the start of what would have been the first treatment phase, so the treatment phase was restarted and that 8-day period was treated as its own separate phase.

Results revealed a reduction in sexual interest in boys and a concurrent shift in sexual interest in same-age male peers, with a shift in predominant sexual interest from boys to same-age peers that began at the start of the B phase (following the unscheduled phase treatment nonadherence) and continued throughout the second baseline and final treatment phase. Whereas the dependent variable did not shift back to baseline levels during the second iteration of phase A, the timing of the commencement of the man's improvements provides evidence for the treatment as a probable cause for the shift, considering it would be difficult for the patient to "unlearn" the skills provided in the treatment phase. Figure 3.3 shows an example of an A-B-A-B design.

B-A-B Design

In the B-A-B design, the treatment (B phase) is applied before a baseline examination (A phase) and the examination ends with a final phase of treatment (B) (Barlow et al., 2009) (see an example of a B-A-B design in Figure 3.4). Many prefer the B-A-B design because active treatment is administered both at the start and at the end of examination. Thus, the B-A-B design offers a clinically indicated experimental strategy for individuals for whom waiting for treatment in order to collect baseline data is contraindicated. Additionally, similar to the A-B-A-B design, the last phase is treatment, a sequence that may increase the likelihood that the patient will

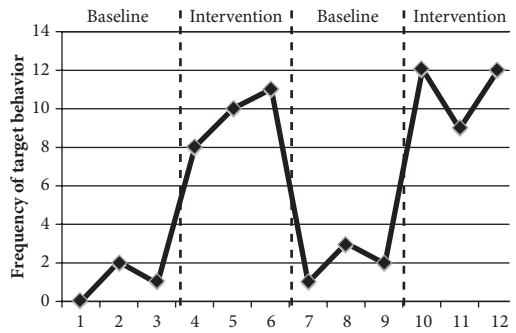


Figure 3.3 Example of an A-B-A-B design.

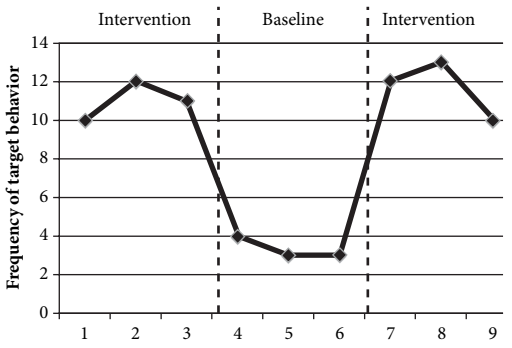


Figure 3.4 Example of a B-A-B design.

continue to benefit from the treatment even after the examination ends.

However, the B-A-B design has limited experimental utility because it is not possible to examine the treatment effects against the natural frequency of the target behavior without a baseline phase occurring before treatment implementation. Although the A phase is marked by the withdrawal of the intervention, no previous baseline has been established in a B-A-B design, prohibiting measurement of the target behavior unaffected by treatment (either during the treatment phase itself or as a remnant of the treatment phase just prior). Thus, the A-B-A-B design is preferable in most cases. An illustration of the problems of a B-A-B design is an early study by Truax and Carkhuff (1965) examining the effects of the Rogerian techniques of empathy and unconditional positive regard on three psychiatric patients' responses in a 1-hour interview. Three 20-minute phases made up the hour-long interview: B, in which the therapist utilized high levels of empathy and unconditional positive regard; A, when these techniques were decreased; and B, when the therapeutic techniques were again increased. Coders who were blind to phase rated and confirmed the relative presence and absence of empathy and unconditional positive regard. The dependent variable of interest was the patient's "intrapersonal exploration." The researchers did identify slightly higher intrapersonal exploration in the B phases relative to the withdrawal (A) phase. However, this investigation does not present a compelling indication of positive intervention effects, as we have no indication of where levels of the target behavior were prior to intervention.

A-B-C-B Design

The A-B-C-B design attempts to control for placebo effects that may affect the dependent variable. To take one example, rather than implementing a return to baseline or withdrawal phase (A), following the initial intervention phase consisting of contingent reinforcement, the amount of reinforcement in the C phase remains the same as in the B phase, but is not contingent on the behavior of the subject (Barlow et al., 2009). For example, if a child received a sticker in phase B for every time he raised his hand to speak in class instead of talking out of turn, in phase C the provision of stickers for the child would not be contingent upon his raising his hand. The C phase thus serves a similar purpose as the placebo phase common in evaluations of pharmaceutical agents.

The principal strength of this design over the traditional A-B-A-B design is that the improvements seen in the B (intervention) phase can be more reliably assigned to the effects of the intervention itself rather than to the effects of participating in an experimental condition. In an A-B-C-B design, the baseline phase cannot be compared against either the B or C phase, as the baseline phase occurs only once and is not repeated for comparison in a later phase of the examination.

One study from the child literature utilized an A-B-C-B design to evaluate a peer-mediated intervention meant to improve social interactions in children with autism (Goldstein, Kaczmarek, Pennington, & Shafer, 1992). Five groups of three children (one with autism, two peers without autism) engaged in coded interactions in which the peers were taught facilitation methods. The A phase consisted of the baseline, with conversation as usual. In the B phase, the two peers were instructed to use the social facilitation strategies they were taught. The C phase consisted of continued use of facilitation strategies, but peers were instructed to use them with the other child instead of with the child with autism. In this phase, they were praised only when they used the newly learned strategies with the peer who did not have autism. The B (peer intervention) phase saw an increase in communicative acts, which returned back to levels similar to baseline in the C phase, and rose again to higher levels during the second B phase, for four of the five children. These outcomes provide evidence for the efficacy of the intervention taught for use during the B phases, above and beyond the effects of simply participating in a study.

Multiple-Baseline Designs

Withdrawal designs are particularly well suited for the evaluation of interventions that would be

less likely to retain effects once they are removed, as is the case in the evaluation of a therapeutic medication with a very short half-life. Some procedures are, however, irreversible (e.g., various surgeries, or the learning of a skill in psychotherapy). How can the clinical researcher evaluate the intervention when it is not possible to completely remove the intervention? In such situations, reversal and withdrawal designs are misguided because withdrawing the intervention may have little effect. When withdrawal or reversal is impossible or unethical, multiple-baseline designs offer a valuable alternative.

Multiple-baseline designs entail applying an intervention to different behaviors, settings, or subjects, while systematically varying the length of the baseline phase for each behavior, setting, or subject (Baer, Wolf, & Risley, 1968). Whereas multiple-baseline designs do not include a withdrawal of treatment, the efficacy of the treatment is demonstrated by reproducing the treatment effects in different behaviors, people, or settings at different times. Accordingly, the multiple-baseline design consists of an A and a B phase, but the A phase is differentially extended for each target behavior, subject, and/or setting. For example, one individual's baseline might last 3 days, another's baseline might last 5 days, and a third individual's baseline might last 7 days. If the intervention is effective, the behavior will not change until the intervention is actually initiated. Thus, analysis in a multiple-baseline design occurs within subjects, settings, or behaviors. Does the behavior change after treatment begins relative to baseline, and among subjects, settings, or behaviors (do other baselines remain stable while one changes)? A strong multiple-baseline strategy has the baseline phase continue until stability of data is observed so that any effects of the intervention can be adequately measured against the stable baseline. Once stability is achieved, the clinical researcher may begin applying the treatment. Another multiple-baseline strategy determines the lengths of baseline intervals *a priori* and then randomly assigns these interval lengths to subjects.

Multiple-baseline designs can take one of three forms: (1) multiple-baseline design across behaviors (example in Fig. 3.5); (2) multiple-baseline design across subjects (example in Fig. 3.6); or (3) multiple-baseline design across settings (Hersen, 1982; Miltenberger, 2001). The *multiple-baseline design across behaviors* examines the effects of an intervention on different behaviors within the same individual. When examining the intervention across behaviors, the clinical researcher applies the intervention in temporal sequence to independent behaviors. Support for an intervention is demonstrated when outcome behaviors improve across the study upon the initiation of treatment targeting those specific behaviors, and not before.

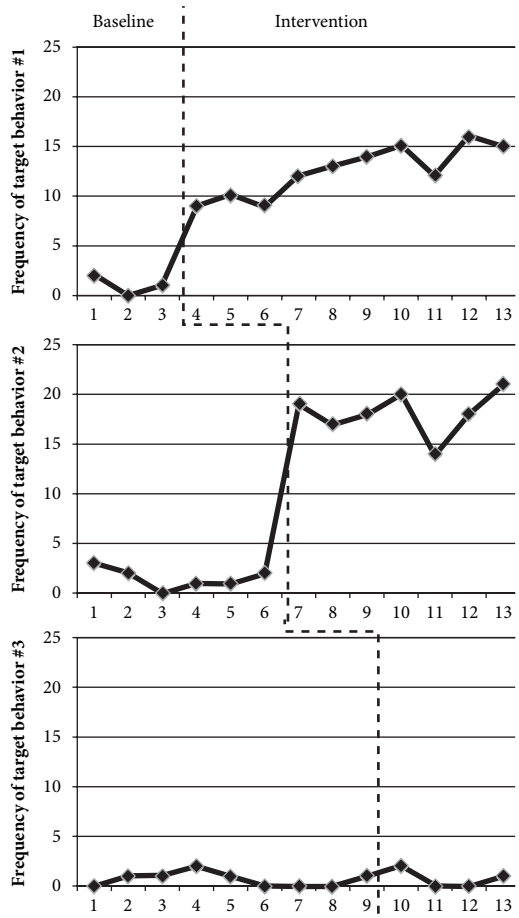


Figure 3.5 Example of multiple-baseline design across behaviors, where behaviors 1 and 2 were specifically targeted by the intervention and behavior 3 was not.

As an example of a multiple-baseline design across behaviors, Lane-Brown and Tate (2010) evaluated a novel treatment for apathy that included positive reinforcement and motivational interviewing in a man with a traumatic brain injury. Specific behaviors targeted were bedroom organization, increasing exercise, and improving social conversations. The first two goals were treated while the latter remained untreated. Lane-Brown and Tate found an increase in goal-directed activity involving organization and exercise after each of these behaviors was targeted by treatment, but no improvement on the untargeted social conversations, providing evidence that it was the treatment that led to observed changes.

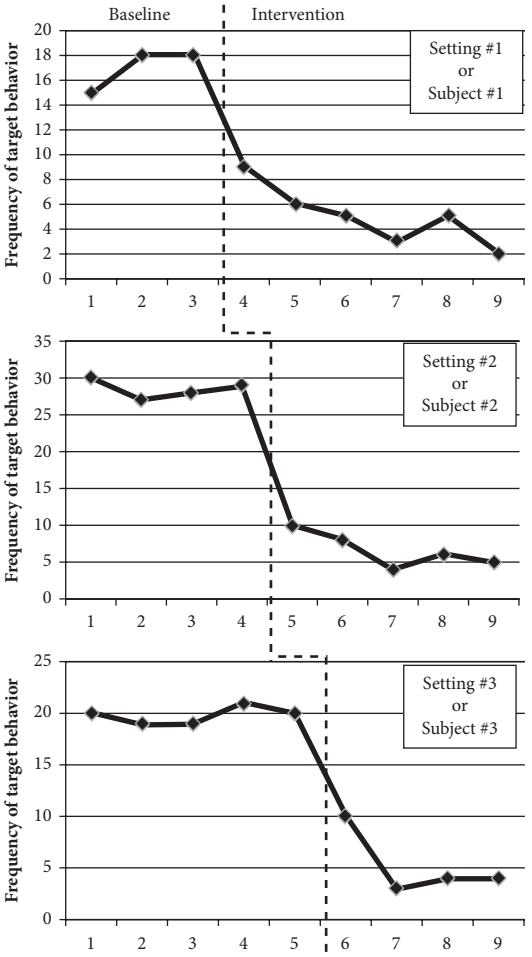


Figure 3.6 Example of multiple-baseline design across settings or subjects.

The *multiple-baseline design across subjects* (or across individuals) examines the effects of intervention on different people with similar presentations, with the duration of the baseline interval varying across subjects. For example, in a study of six individuals, two may undergo a 2-week baseline interval prior to treatment, two may undergo a 4-week baseline interval prior to treatment, and two may undergo a 6-week baseline interval prior to treatment. The effect of an intervention is demonstrated when a change in each person's functioning is obtained after the initiation of treatment, and not before.

Choate, Pincus, Eyberg, and Barlow (2005) utilized the multiple-baseline design across subjects to examine an adaptation of Parent-Child Interaction Therapy (PCIT) to treat early separation anxiety disorder. Treatment of three different children was implemented after 1, 2, and 4 weeks of baseline monitoring of anxiety symptoms. PCIT consists of two stages: the child directed interaction (CDI) phase, during which parents are taught to follow the lead of the child, and the parent directed interaction (PDI) phase, during which parents learn to effectively direct and lead the child (Herschell, Calzada, Eyberg, & McNeil, 2002). Anxiety symptoms remained stable during the baseline period for all three subjects and began decreasing only after the initiation of treatment, particularly during the PDI portion of treatment, showing preliminary support for the use of adapted PCIT to treat preschoolers with separation anxiety disorder.

In a *multiple-baseline design across settings*, treatment is applied in sequence across new and different settings (such as at home, at school, and with peers) (Freeman, 2003). These designs demonstrate treatment efficacy when changes occur in each setting when, and only when, the intervention is implemented in that setting. Kay, Harchik, and Luiselli (2006) used such a design to evaluate a multicomponent behavior intervention using compensatory responses and positive reinforcement to reduce drooling in a 17-year-old boy with autism. The intervention was introduced after varying numbers of days in three settings (the classroom, the community, and cooking class), with decreased drooling occurring in each setting only after the intervention was introduced in that setting.

Kazdin and Kopel (1975) provide recommendations for how to be sure that the treatment is affecting the target variable. Specifically, the baselines should be as different as possible from each other in length, at least four baselines should be used, and/or treatment should be withdrawn and reapplied if necessary to demonstrate that the treatment causes the change in the target variable. The number of baselines that are needed has been deliberated in the literature, with a consensus that three or four baselines are necessary to be sure that observed changes are the result of the treatment (Barlow et al., 2009; Kazdin & Kopel, 1975; Wolf & Risley, 1971).

The strength of the multiple-baseline design comes largely from the ability to demonstrate the efficacy of an intervention by showing that the desired change occurs only when the intervention is applied to the behavior, subject, or setting specifically targeted (Barlow et al., 2009). One of the biggest advantages of the multiple-baseline design is that it allows for multiple behaviors to be examined at one time, which is more similar to naturalistic situations, and allows the behaviors to be measured in the context of each other (Barlow et al., 2009)—for example, in the case of comorbid conditions.

However, unlike withdrawal designs, multiple-baseline designs control only the introduction, but not the removal, of treatment. Thus, when appropriate, withdrawal designs are able to yield more compelling evidence for causal conclusions. Additionally, the multiple-baseline design's strength decreases if fewer than three or four settings, behaviors, or individuals are measured. Finally, there are limitations to the multiple-baseline design involving generalization, but the possibility for generalization can be further evaluated utilizing "generalization tests" (see Kendall, 1981).

Moving from Idiographic to Nomothetic Group Design Evaluations

Whereas single-case experimental designs and multiple-baseline series inform our understanding of individual behavior change and play key roles in treatment development, nomothetic group experimental designs are essential for establishing treatment efficacy and effectiveness, and for meaningfully influencing health care policy and practice. Specifically, adequately powered RCTs that maximize both scientific rigor and clinical relevance constitute the field's "gold standard" research design for establishing broad empirical support for a treatment (Chambless & Hollon, 1998; Kendall & Comer, 2011). Such work entails a well-defined independent variable (i.e., manualized treatment protocols), appropriate control condition(s), a comprehensive multimodal/multi-informant assessment strategy, treatment fidelity checks, statistical and clinical significance testing, evaluation of response across time, and an adequately powered sample of clinically representative participants to enable statistical judgments that are both reliable and valid (see Kendall, Comer, & Chow, this volume, for a full consideration of RCT methods and design). Needless to say, such undertakings are enormously time- and resource-intensive, and so entering into a large-scale RCT first requires careful preparation to minimize the risk of a failed study, unfortunate wasting of time and resources, and unwarranted burden on study participants. Prior to conducting a large adequately powered RCT, a small pilot RCT is warranted.

Appropriate Use and Design of the Small Pilot RCT

The empirical preparation for a large-scale RCT is the purview of the *small pilot RCT*. Many erroneously perceive the sole function of the small pilot RCT as providing preliminary information on the feasibility and acceptability of the experimental treatment, or providing a preliminary indication of the effectiveness of the experimental treatment. Too often researchers fail to appreciate the small pilot RCT's more fundamental role in providing preliminary information on the feasibility and acceptability of the *study design* to be used in the subsequent large-scale treatment evaluation. The pilot RCT serves as a check on the research team's ability to recruit, treat, and retain participants across randomization and key study points (e.g., as a check on the availability of eligible and willing participants using the proposed recruitment methods, to test the feasibility of assessment and treatment protocols, to evaluate whether the study protocol sufficiently retains target participants across randomization or whether participants systematically drop out when assigned to a less-preferred treatment condition, to evaluate whether participant compensation is sufficient to recruit participants to complete assessments long after treatment has been completed, etc.). The small pilot RCT thus provides researchers an opportunity to identify and correct potential "glitches" in the research design prior to the funding and initiation of an adequately powered large-scale RCT (Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006), and accordingly the pilot RCT should ideally implement an identical design to that foreseen for the subsequent large-scale RCT.

Failure to appreciate this fundamental role of the small pilot trial as a check on the study design can have dramatic effects on the design of a pilot trial, which can in turn have unfortunate consequences for a program of research. Consider the following cautionary example of a researcher who misguidedly perceives the sole function of pilot work as providing preliminary information on the feasibility and acceptability of the experimental treatment:

> A researcher spends considerable efforts conceptualizing and developing an intervention for a target clinical population based on theory, empirical research, and extensive consultation with noted experts in the area. The researcher appreciates the need for an adequately powered RCT in the establishment of empirical support for his treatment, appreciates that such an endeavor will require considerable funding, and also appreciates that a grant review committee will require pilot data before it would consider recommending funding for a proposed large-scale RCT. And so the researcher secures small internal funding to pilot test his treatment, and calculates that with this money

he is able to treat and evaluate 16 participants across pretreatment, posttreatment, and 6-month follow-up.

Given the researcher's limited funds for the pilot work, and given his misguided sole focus on establishing the feasibility and acceptability of his novel treatment with the pilot data, he decides to run all of the pilot subjects through the experimental treatment. "After all," the researcher thinks to himself, "since a pilot study is underpowered to statistically test outcomes against a control condition, I might as well run as many subjects as I can through my new treatment so that I can have all the more data on treatment credibility and consumer satisfaction." The researcher further decides that since pilot work is by design underpowered to enable statistical judgments about treatment efficacy, to save costs he would rely solely on self-reports rather than on lengthy structured diagnostic interviews, although he does plan to include diagnostic interviews in the subsequent large-scale RCT design. Finally, the researcher calculates that if he cuts the 6-month follow-up assessments from the pilot design, he can run four more subjects through the experimental treatment. At the end of the pilot trial, he treats 20 subjects with the experimental treatment (with only three dropouts) and collects consumer satisfaction forms providing preliminary indication of treatment feasibility and acceptability.

The researcher includes these encouraging pilot data in a well-written grant submission to fund an impressively designed large-scale RCT comparing his experimental treatment to a credible education/support/attention control condition with a 6-month follow-up, but is surprised when the scientific review committee recommends against funding his work due to "inadequate pilot testing." The summary statements from the review note that the researcher does not provide any evidence that he can recruit and retain subjects across a randomization, or that his team can deliver the education/support/attention control condition proposed, or that subjects randomly assigned to this control condition will not drop out when they learn of their treatment assignment. The committee also questions whether his team is sufficiently trained to conduct the diagnostic interviews proposed in the study protocol, as these were not included in the pilot trial. Because there were no 6-month follow-up evaluations included in the pilot work, the committee expresses uncertainty about the researcher's ability to compel participants to return for assessments so long after treatment has completed, and wonder whether his proposed $10 compensation for participating in assessments is sufficient to maximize participation.

In the above example, the researcher's failure to appreciate the role of pilot work in gathering preliminary information on the feasibility and acceptability of proposed study procedures—and not solely the feasibility and acceptability of the treatment itself—interfered with his ability to secure funding for an adequately powered and controlled evaluation of his treatment.

This researcher would have been better off using the pilot funding to conduct a small pilot RCT implementing an identical design foreseen for the subsequent large-scale RCT. Specifically, a pilot design that randomized 16 participants across the two treatments that he intended to include in the large-scale RCT and employed diagnostic interviews and 6-month follow-up assessments would have provided more compelling evidence that his proposed large-scale RCT was worth the sizable requested investment. Given that small pilot samples are not sufficiently powered to enable stable efficacy judgments (Kraemer et al., 2006), the additional 12 subjects he gained by excluding a control group and abandoning diagnostic and follow-up assessments, in truth, provided no incremental support for his treatment. Instead, the inadequate pilot design left the review committee with too many questions about his team's ability to implement and retain subjects across a randomization procedure, the team's ability to implement the education control treatment faithfully without inadvertently including elements of the experimental treatment, the team's ability to adequately conduct diagnostic interviews, and the team's ability to retain subjects across the proposed long-term follow-up. Researchers who receive this type of feedback from grant review committees are undoubtedly disappointed, but not nearly as disappointed as those researchers who have invested several years and considerable resources into a controlled trial only to realize midway through the study that key glitches in their study design that could have been easily averted are systematically interfering with the ability to truly evaluate treatment efficacy or to meaningfully interpret the data.

Caution Concerning the Misuse of Pilot Data for the Purposes of Power Calculations

It is critical to caution researchers against the common misuse of data drawn from small pilot studies for the purposes of power calculations in the design of a subsequent large-scale RCT. As well articulated elsewhere (Cohen, 1988; Kraemer &

Thiemann, 1987; see the chapter by Kraemer in this volume), *power* refers to the probability of accurately rejecting a null hypothesis (e.g., the effect of an experimental treatment is comparable to the effect of a control treatment) when the null hypothesis is indeed untrue. Designing an adequately powered RCT study entails recruiting a sample large enough to yield reliably different treatment response scores across conditions if true group response differences do indeed exist. Conventional calculations call for the researcher to determine the needed sample size via calculations that consider an *expected effect size* (in RCT data, typically the magnitude of difference in treatment response across groups) in the context of an acceptably low α level (i.e., the probabilty of rejecting the null hypothesis if it is indeed true; consensus typically stipulates $\alpha \leq .05$) and an acceptably high level of power (consensus typically stipulates power $\geq .80$) (which sets the probability of correctly rejecting the null hypothesis when there is a true effect at four in five tests).

Although conventions stipulate acceptable α and power levels to incorporate into sample size calculations, broad conventions do not stipulate an expected effect size magnitude to include because this will vary widely across diverse clinical populations and across varied treatments. Whereas an exposure-based treatment for specific phobias may expectedly yield a relatively large effect size, a bibliotherapy treatment for borderline personality disorder may expectedly yield a very small effect size. To estimate an expected effect size for the design of an adequately powered study, the researcher must rely on theory regarding the specific clinical population and the treatment being evaluated, as well as the magnitude of effects found in related studies. Indeed, expert guidelines argue that rationale and justification for a proposed hypothesis-testing study should be drawn "from previous research" (Wilkinson & Task Force on Statistical Inference, 1999).

Commonly, researchers will accordingly use data from their pilot RCT to estimate an expected effect size for a proposed large-scale RCT. For example, if a small pilot RCT ($n = 15$) identified a large treatment effect (e.g., $d = 0.8$), a researcher might use this effect size to guide power calculations for determining the necessary sample size for a proposed large-scale RCT. But as Kraemer and colleagues (2006) mathematically demonstrate, this misguided practice can lead to the design of underpowered studies positioned to retain the null hypothesis when in fact true treatment differences exist, or to a failure to pursue large-scale research that would identify meaningful treatment effects. Because a limited sample size can yield large variability in effects, effect sizes drawn from underpowered studies (such as small pilot studies) result in effect size estimates that are unstable. In the above example, although a large treatment effect was found in the pilot trial, the true treatment effect may in fact be moderate but meaningful (e.g., $d = 0.5$). As a larger sample size is required to reliably detect a moderate effect versus a large effect, a study designed to simply capture a large effect is at increased risk to retain the null hypothesis when in fact there are true treatment differences (i.e., a power analysis based on a predicted large effect would estimate the need for a smaller sample than would one based on a predicted moderate effect). In this scenario, after a thorough time- and resource-intensive RCT, the researcher would erroneously conclude that his treatment does not "work." Accordingly, the researcher is better justified to rely on related work in the literature using adequately powered samples to evaluate the effect of similar treatment methods for neighboring clinical conditions than to rely on underpowered pilot work, even though the pilot work examined the very treatment for the very condition under question.

Discussion

The past 25 years have witnessed tremendous progress in the advancement of evidence-based practice. Many contemporary treatment guidelines (e.g., Chambless & Hollon, 1998; Silverman & Hinshaw, 2008) appropriately privilege the outcomes of large randomized group comparison trials over other research methodologies in the identification of empirically supported treatments. Large RCTs are undoubtedly the most rigorous and experimentally controlled methodology we have with which to inform broad public policy decisions and mental health care debates. However, key limitations in the generality of obtained results highlight the additional need for data drawn from complementary methods that add to the rigorous evidence yielded by the RCT. Indeed, consensus guidelines for evidence-based practice explicitly call for supporting evidence drawn from a broad portfolio of research methods and strategies, each with its own advantages and limitations.

The multiple strengths of single-case experimental designs and small pilot trials should place these designs firmly in the comprehensive portfolio of informative designs for evaluating evidence-based practices in mental health care. Regrettably, the

prominence of and appreciation for such designs has waned over the past several decades, as evidenced by their declining representation in leading clinical psychology journals and by the limited availability of funding granted to work utilizing these designs. It may be that for many, deliberation on the merits of single-case designs mistakenly lumps these designs with unrelated methods that also rely on small *n* samples but do not incorporate experimental manipulation of treatment initiation and discontinuation or systematic observation of target behaviors (e.g., case histories, case series, retrospective chart reviews). Importantly, low statistical power and low experimental control are distinct methodological constructs.

In many ways, comprehensive treatment evaluation must begin and end with the study of change in the individual. The large-scale RCT cannot be conducted in a vacuum. Systematic research activities focusing on individual treatment-related changes are needed in preparation for large clinical trials, and systematic research activities evaluating the successful application of RCT-supported treatments to individuals that may vary in important and meaningful ways are needed before attempting large-scale implementation of supported treatments in practice settings. Too often, researchers rush through the treatment development stage in order to focus their efforts on randomized controlled outcomes. Often the result is a very rigorous evaluation of a treatment package that could have been substantially improved had the developers invested the time in single-case design research activities to evaluate the "how," "why," and "when" of treatment-related change. In addition, too often time- and resource-intensive RCTs fail to recruit and retain participants across study procedures when a relatively inexpensive pilot trial could have corrected simple unforeseeable glitches in the research design. And too often treatments supported in tightly controlled RCTs are expected to be applied in broad mental health care settings without first evaluating the supported treatment's effects in individual participants across practice settings. In such cases, the result can be a misguided attempt to shoehorn treatment strategies shown to be effective in controlled laboratory settings into practice settings that may differ in important and meaningful ways—an effort that can inadvertently increase the already regrettable wedge between research and practice communities.

We hope this chapter has provided a renewed sense of urgency for the role of single-case designs, including multiple-baseline and withdrawal designs, and small pilot trial designs in the evaluation of evidence-based practice for mental health conditions. Alongside RCTs, small experimental designs play a crucial role in developing interventions and testing their clinical utility among groups of people and among individuals. Moreover, small experimental designs are easily adaptable in clinical settings and laboratories alike, affording clinicians, many of whom have limited resources to conduct large-scale research studies, a greater opportunity to contribute to the growing literature on evidence-based psychological treatments. Increased utilization of single-case, multiple-baseline, and small pilot trial designs would significantly enhance our understanding of the effects of mental health treatments and would more efficiently elucidate what treatments work best for whom. Appropriately designed small experimental designs are an important and necessary component of the evaluation of any psychological treatment, and increasing their frequency in the research literature will significantly enhance the understanding about the benefits and weaknesses of evidence-based psychological treatments.

References

American Psychological Association. (2002). Criteria for practice guideline development and evaluation. *American Psychologist*, *57*(12), 1048–1051. doi: 10.1037/0003-066x.57.12.1048

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*(1), 91–97. doi: 10.1901/jaba.1968.1-91

Bandura, A. (1969). *Principles of behavior modification*. New York: Holt, Rinehart and Winston.

Barlow, D. H., & Hersen, M. (1973). Single-case experimental designs: Uses in applied clinical research. *Archives of General Psychiatry*, *29*(3), 319–325.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*(1), 19–21. doi: 10.1111/j.1745-6924.2009.01088.x

Barlow, D. H., Nock, M., & Hersen, M. (2009). *Single case experimental designs: strategies for studying behavior change* (3rd ed.). Boston: Pearson/Allyn and Bacon.

Barron, F., & Leary, T. F. (1955). Changes in psychoneurotic patients with and without psychotherapy. *Journal of Consulting Psychology*, *19*(4), 239–245. doi:10.1037/h0044784

Bolger, H. (1965). The case study method. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 28–39). New York: McGraw-Hill.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409–429. doi: 10.1037/h0027982

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: R. McNally.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18. doi: 10.1037/0022-006x.66.1.7

Choate, M. L., Pincus, D. B., Eyberg, S. M., & Barlow, D. H. (2005). Parent-Child Interaction Therapy for treatment of separation anxiety disorder in young children: A pilot study. *Cognitive and Behavioral Practice*, *12*(1), 126–135. doi: 10.1016/s1077-7229(05)80047-1

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, *12*(4), 425–434. doi: 10.1177/014662168801200410

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cooper, M., Todd, G., Turner, H., & Wells, A. (2007). Cognitive therapy for bulimia nervosa: An A-B replication series. *Clinical Psychology & Psychotherapy*, *14*(5), 402–411. doi: 10.1002/cpp.548

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. doi: 10.1037/h0047358

Eysenck, H. J. (1952). The effects of psychotherapy: an evaluation. *Journal of Consulting Psychology*, *16*(5), 319–324. doi: 10.1037/h0063633

Freeman, K. A. (2003). Single subject designs. In J. C. Thomas & M. Hersen (Eds.), *Understanding research in clinical and counseling psychology* (pp. 181–208). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Goldstein, H., Kaczmarek, L., Pennington, R., & Shafer, K. (1992). Peer-mediated intervention: Attending to, commenting on, and acknowledging the behavior of preschoolers with autism. *Journal of Applied Behavior Analysis*, *25*(2), 289–305. doi: 10.1901/jaba.1992. 25–289

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist practitioner: research and accountability in the age of managed care*. Needham, MA: Allyn and Bacon.

Herschell, A. D., Calzada, E. J., Eyberg, S. M., & McNeil, C. B. (2002). Clinical issues in parent-child interaction therapy. *Cognitive and Behavioral Practice*, *9*(1), 16–27. doi: 10.1016/s1077-7229(02)80035-9

Hersen, M. (1982). Single-case experimental designs. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 167–203). New York: Plenum Press.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: strategies for studying behavior change* (1st ed., Vol. 56). New York: Pergamon Press.

Hunter, J. A., Ram, N., & Ryback, R. (2008). Use of satiation therapy in the treatment of adolescent-manifest sexual interest in male children: A single-case, repeated measures design. *Clinical Case Studies*, *7*(1), 54–74. doi: 10.1177/1534650107304773

Kay, S., Harchik, A. F., & Luiselli, J. K. (2006). Elimination of drooling by an adolescent student with autism attending public high school. *Journal of Positive Behavior Interventions*, *8*(1), 24–28. doi: 10.1177/10983007060080010401

Kazdin, A. E. (1973). Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, *6*(3), 517–531. doi: 10.1901/jaba.1973.6–517

Kazdin, A. E. (1982). *Single-case research designs: methods for clinical and applied settings*. New York: Oxford University Press.

Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth/Thompson Learning.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn and Bacon.

Kazdin, A. E., & Kopel, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, *6*(5), 601–608. doi: 10.1016/s0005-7894(75)80181-x

Kendall, P. C. (1981). Assessing generalization and the single-subject strategies. *Behavior Modification*, *5*(3), 307–319. doi:10.1177/014544558153001

Kendall, P. C., & Comer, J. S. (2011). Research methods in clinical psychology. In D. H. Barlow (Ed.), *The Oxford handbook of clinical psychology* (pp. 52–75). New York: Oxford University Press.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, *63*(5), 484–489. doi: 10.1001/archpsyc.63.5.484

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Thousand Oaks, CA: Sage Publications, Inc.

Landsverk, J., Brown, C., Rolls Reutz, J., Palinkas, L., & Horwitz, S. (2011). Design elements in implementation research: A structured review of child welfare and child mental health studies. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 54–63. doi: 10.1007/s10488-010-0315-y

Lane-Brown, A. P., & Tate, R. P. (2010). Evaluation of an intervention for apathy after traumatic brain injury: A multiple-baseline, single-case experimental design. *Journal of Head Trauma Rehabilitation*, *25*(6), 459–469.

March, J. S., Silva, S. G., Compton, S., Shapiro, M., Califf, R., & Krishnan, R. (2005). The case for practical clinical trials in psychiatry. *American Journal of Psychiatry*, *162*(5), 836–846. doi: 10.1176/appi.ajp.162.5.836

Marshall, W. L. (1979). Satiation therapy: A procedure for reducing deviant sexual arousal. *Journal of Applied Behavior Analysis*, *12*(3), 377–389. doi: 10.1901/jaba.1979.12–377

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, *5*(1), 87–101. doi: 10.1037/1082-989x.5.1.87

Miltenberger, R. G. (2001). *Behavior modification: Principles and procedures* (2nd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Moore, T. R., Gilles, E., McComas, J. J., & Symons, F. J. (2010). Functional analysis and treatment of self-injurious behaviour in a young child with traumatic brain injury. *Brain Injury, 24*(12), 1511–1518. doi: 10.3109/02699052.2010.523043

Nock, M. K., Goldman, J. L., Wang, Y., & Albano, A. M. (2004). From science to practice: The flexible use of evidence-based treatments in clinical settings. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(6), 777–780. doi: 10.1097/01.chi.0000120023.14101.58

Nock, M. K., & Kurtz, S. M. S. (2005). Direct behavioral observation in school settings: Bringing science to practice. *Cognitive and Behavioral Practice*, *12*(3), 359–370. doi: 10.1016/s1077-7229(05)80058-6

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. doi: 10.1037/h0043424

Powers, E., & Witmer, H. (1951). An experiment in the prevention of delinquency; the Cambridge-Somerville Youth Study. New York: Columbia University Press.

Risley, T. R., & Wolf, M. M. (1972). Strategies for analyzing behavioral change over time. In J. Nesselroade & H. Reese (Eds.), *Life-span developmental psychology. Methodological issues* (pp. 175–183). New York: Academic Press.

Rounsaville, B. J., Carroll, K. M., & Onken, L. S. (2001). A stage model of behavioral therapies research: Getting started and moving on from stage I. *Clinical Psychology: Science and Practice*, *8*(2), 133–142. doi: 10.1093/clipsy/8.2.133

Sidman, M. (1960). *Tactics of scientific research*. Oxford, England: Basic Books.

Silverman, W. K., & Hinshaw, S. P. (2008). The second special issue on evidence-based psychosocial treatments for children and adolescents: A 10-year update. *Journal of Clinical Child and Adolescent Psychology*, *37*(1), 1–7. doi: 10.1080/15374410701817725

Suveg, C., Comer, J. S., Furr, J. M., & Kendall, P. C. (2006). Adapting manualized CBT for a cognitively delayed child with multiple anxiety disorders. *Clinical Case Studies*, *5*(6), 488–510. doi: 10.1177/1534650106290371

Truax, C. B., & Carkhuff, R. R. (1965). Experimental manipulation of therapeutic conditions. *Journal of Consulting Psychology*, *29*(2), 119–124. doi: 10.1037/h0021927

Wallenstein, M. B., & Nock, M. K. (2007). Physical exercise as a treatment for non-suicidal self-injury: Evidence from a single case study. *American Journal of Psychiatry*, *164*(2), 350–351. doi: 10.1176/appi.ajp.164.2.350-a

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, *3*(1), 1–14. doi: 10.1037/h0069608

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. doi: 10.1037/0003-066x.54.8.594

Wolf, M. M., & Risley, T. R. (1971). Reinforcement: Applied research. In R. Glaser (Ed.), *The nature of reinforcement* (pp. 310–325). New York: Academic Press.

CHAPTER

4 The Randomized Controlled Trial: Basics and Beyond

Philip C. Kendall, Jonathan S. Comer, *and* Candice Chow

Abstract

This chapter describes methodological and design considerations central to the scientific evaluation of clinical treatment methods via randomized clinical trials (RCTs). Matters of design, procedure, measurement, data analysis, and reporting are each considered in turn. Specifically, the authors examine different types of controlled comparisons, random assignment, the evaluation of treatment response across time, participant selection, study setting, properly defining and checking the integrity of the independent variable (i.e., treatment condition), dealing with participant attrition and missing data, evaluating clinical significance and mechanisms of change, and consolidated standards for communicating study findings to the scientific community. After addressing considerations related to the design and implementation of the traditional RCT, the authors turn their attention to important extensions and variations of the RCT. These treatment study designs include equivalency designs, sequenced treatment designs, prescriptive designs, adaptive designs, and preferential treatment designs. Examples from the recent clinical psychology literature are provided, and guidelines are suggested for conducting treatment evaluations that maximize both scientific rigor and clinical relevance.

Key Words: Randomized clinical trial, RCT, normative comparisons, random assignment, treatment integrity, equivalency designs, sequenced treatment designs

The randomized controlled trial (RCT)—a group comparison design in which participants are randomly assigned to treatment conditions—constitutes the most rigorous and objective methodological design for evaluating therapeutic outcomes. In this chapter we focus on RCT research strategies that maximize both scientific rigor and clinical relevance (for consideration of single-case, multiple-baseline, and small pilot trial designs, see Chapter 3 in this volume). We organize the present chapter around (a) RCT design considerations, (b) RCT procedural considerations, (c) RCT measurement considerations, (d) RCT data analysis, and (e) RCT reporting. We then turn our attention to extensions and variations of the traditional RCT, which offer various adjustments for clinical generalizability, while at the same time sacrificing important elements of internal validity. Although all of the methodological and design ideals presented may not always be achieved within a single RCT, our discussions provide exemplars of the RCT.

Design Considerations

To adequately assess the *causal* impact of a therapeutic intervention, clinical researchers must use control procedures derived from experimental science. In the RCT, the intervention applied constitutes the experimental manipulation, and thus to have confidence that an intervention is responsible for observed changes, extraneous factors must be experimentally "controlled." The objective is to distinguish intervention effects from any changes that

result from other factors, such as the passage of time, patient expectancies of change, therapist attention, repeated assessments, and simple regression to the mean. To maximize internal validity, the clinical researcher must carefully select control/comparison condition(s), randomly assign participants across treatment conditions, and systematically evaluate treatment response across time. We now consider each of these RCT research strategies in turn.

Selecting Control Condition(s)

Comparisons of participants randomly assigned to different treatment conditions are essential to control for factors other than treatment. In a "controlled" treatment evaluation, comparable participants are randomly placed into either the *experimental condition*, composed of those who receive the intervention, or a *control condition*, composed of those who do not receive the intervention. The efficacy of treatment over and above the outcome produced by extraneous factors (e.g., the passage of time) can be determined by comparing prospective changes shown by participants across conditions.

Importantly, not all control conditions are "created equal." Deciding which form of control condition to select for a particular study (e.g., no treatment, waitlist, attention-placebo, standard treatment as usual) requires careful deliberation (see Table 4.1 for recent examples from the literature). In a *no-treatment control condition*, comparison participants are evaluated in repeated assessments, separated by an interval of time equal in duration to the treatment provided to those in the experimental treatment condition. Any changes seen in the treated participants are compared to changes seen in the nontreated participants. When, relative to nontreated participants, the treated participants show significantly greater improvements, the experimental treatment may be credited with producing the observed changes. Several important rival hypotheses are eliminated in a no-treatment design, including effects due to the passage of time, maturation,

Table 4.1 Types of Control Conditions in Treatment Outcome Research

		Recent Example in Literature	
Control Condition	**Definition**	**Description**	**Reference**
No-treatment control	Control participants are administered assessments on repeated occasions, separated by an interval of time equal to the length of treatment.	Adults with anxiety symptoms were randomly assigned to a standard self-help condition, an augmented self-help condition, or a control condition in which they did not receive any intervention.	Varley et al. (2011)
Waitlist control	Control participants are assessed before and after a designated duration of time but receive the treatment following the waiting period. They may anticipate change due to therapy.	Adolescents with anxiety disorders were randomly assigned to Internet-delivered CBT, face-to-face CBT, or to a waitlist control group.	Spence et al. (2011)
Attention-placebo/ nonspecific control	Control participants receive a treatment that involves nonspecific factors (e.g., attention, contact with a therapist).	School-age children with anxiety symptoms were randomly assigned to either a cognitive-behavioral group intervention or an attention control in which students were read to in small groups.	Miller et al. (2011)
Standard treatment/ routine care control	Control participants receive an intervention that is the current practice for treatment of the problem under study.	Depressed veterans were randomly assigned to either telephone-administered cognitive-behavioral therapy or standard care through community-based outpatient clinics.	Mohr et al. (2011)

spontaneous remission, and regression to the mean. Importantly, however, other potentially important confounding factors not specific to the experimental treatment—such as patient expectancies to get better, or meeting with a caring and attentive clinician—are not ruled out in a no-treatment control design. Accordingly, no-treatment control conditions may be useful in earlier stages of treatment development, but to establish broad empirical support for an intervention, more informative control procedures are preferred.

A more revealing variant of the no-treatment condition is the *waitlist condition*. Here, participants in the waitlist condition expect that after a certain period of time they will receive treatment, and accordingly may anticipate upcoming changes (which may in turn affect their symptoms). Changes are evaluated at uniform intervals across the waitlist and experimental conditions, and if we assume the participants in the waitlist and treatment conditions are comparable (e.g., comparable baseline symptom severity and gender, age, and ethnicity distributions), we can then infer that changes in the treated participants relative to waitlist participants are likely due to the intervention rather than to expectations of impending change. However, as with no-treatment conditions, waitlist conditions are of limited value for evaluating treatments that have already been examined relative to "inactive" conditions.

No-treatment and waitlist conditions in study designs introduce important ethical considerations, particularly with vulnerable populations (see Kendall & Suveg, 2008). For ethical purposes, the functioning of waitlist participants must be carefully monitored to ensure that they are safely able to tolerate the treatment delay. If a waitlist participant experiences a clinical emergency requiring immediate professional attention during the waitlist interval, the provision of emergency professional services undoubtedly compromises the integrity of the waitlist condition. In addition, to maximize internal validity, the duration of the control condition should be equal to the duration of the experimental treatment condition to ensure that differences in response across conditions cannot be attributed simply to differential passages of time. Now suppose a 24-session treatment takes 6 months to provide—is it ethical to withhold treatment for such a long wait period (see Bersoff & Bersoff, 1999)? The ethical response to this question varies across clinical conditions. It may be ethical to incorporate a waitlist design when evaluating an experimental treatment for obesity, but a waitlist design may be unethical when evaluating an experimental treatment for suicidal participants. Moreover, with increasing waitlist durations, the problem of differential attrition arises, which compromises study interpretation. If attrition rates are higher in a waitlist condition, the sample in the control condition may be different from the sample in the treatment condition, and no longer representative of the larger group. For appropriate interpretation of study results, it is important to recognize that the smaller waitlist group at the end of the study now represents only patients who could tolerate and withstand a prolonged waitlist period.

An alternative to waitlist control condition is the *attention-placebo control condition* (or nonspecific treatment condition), which accounts for key effects that might be due simply to regularly meeting with and getting the attention of a warm and knowledgeable therapist. For example, in a recent RCT, Kendall and colleagues (2008) randomly assigned children with anxiety disorders to receive one of two forms of cognitive-behavioral treatment (CBT; either individual or family CBT) or to a manual-based family education, support, and attention (FESA) condition. Individual and family-based CBT showed superiority over FESA in reducing children's principal anxiety diagnosis. Given the attentive and supportive nature of FESA, it could be inferred that gains associated with CBT were not likely attributable to "common therapy factors" such as learning about emotions, receiving support from an understanding therapist, and having opportunities to discuss the child's difficulties.

Developing and implementing a successful attention-placebo control condition requires careful deliberation. Attention placebos must credibly instill positive expectations in participants and provide comparable professional contact, while at the same time they must be devoid of specific therapeutic techniques hypothesized to be effective. For ethical purposes, participants must be fully informed of and willing to take a chance on receiving a psychosocial placebo condition. Even then, a credible attention-placebo condition may be difficult for therapists to accomplish, particularly if they do not believe that the treatment will offer any benefit to the participant. Methodologically, it is difficult to ensure that study therapists share comparable positive expectancies for their attention-placebo participants as they do for their participants who are receiving more active treatment (O'Leary & Borkovec, 1978). "Demand characteristics" suggest that when study therapists predict a favorable treatment response, participants will tend to improve accordingly (Kazdin, 2003),

which in turn affects the interpretability of study findings. Similarly, whereas participants in an attention-placebo condition may have high baseline expectations, they may grow disenchanted when no meaningful changes are emerging. The clinical researcher is wise to assess participant expectations for change across conditions so that if an experimental treatment outperforms an attention-placebo control condition, the impact of differential participant expectations across conditions can be evaluated.

Inclusion of an attention-placebo control condition, when carefully designed, offers advantages from an internal validity standpoint. Treatment components across conditions are carefully specified and the clinical researcher maintains tight control over the differential experiences of participants across conditions. At the same time, such designs typically compare an experimental treatment to a treatment condition that has been developed for the purposes of the study and that does not exist in actual clinical practice. The use of a *standard treatment comparison condition* (or treatment-as-usual condition) affords evaluation of an experimental treatment relative to the intervention that is currently available and being applied. Including standard treatment as the control condition offers advantages over attention-placebo, waitlist, and no-treatment controls. Ethical concerns about no-treatment conditions are quelled, and, as all participants receive care, attrition is likely to be minimized, and nonspecific factors are likely to be equated (Kazdin, 2003). When the experimental treatment and the standard care intervention share comparable durations and participant and therapist expectancies, the researcher can evaluate the relative efficacy of the interventions.

In a recent example, Mufson and colleagues (2004) randomly assigned depressed adolescents to interpersonal psychotherapy (IPT-A) or to "treatment as usual" in school-based mental health clinics. Adolescents treated with IPT-A relative to treatment as usual showed greater symptom reduction and improved overall functioning. Given this design, the researchers were able to infer that IPT-A outperformed the existing standard of care for depressed adolescents in the school settings. Importantly, in standard treatment comparisons, it is critical that both the experimental treatment and the standard (routine) treatment are implemented in a high-quality fashion (Kendall & Hollon, 1983).

Random Assignment

To achieve baseline comparability between study conditions, *random assignment* is essential. Random assignment in the context of an RCT ensures that every participant has an equal chance of being assigned to the active treatment condition or to the control condition(s). Random assignment, however, does not guarantee comparability across conditions—simply as a result of chance, one resultant group may be different on some variables (e.g., household income, occupational impairment, comorbidity). Appropriate statistical tests can be used to evaluate the comparability of participants across treatment conditions.

Problems arise when random assignment is not incorporated into a group-comparison design of treatment response. Consider a situation in which participants do not have an equal chance of being assigned to the experimental and control conditions. For example, suppose a researcher were to allow depressed participants to elect for themselves whether to participate in the active treatment or in a waitlist treatment condition. If participants in the active treatment condition subsequently showed greater symptom reductions than waitlist participants, the research cannot rule out the possibility that post-treatment symptom differences could have resulted from prestudy differences between the participants (e.g., selection bias). Participants who choose not to receive treatment immediately may not be ready to work on their depression and may be meaningfully different from those depressed participants who are immediately ready to work on their symptoms.

Although random assignment does not ensure participant comparability across conditions on all measures, randomization procedures do rigorously maximize the likelihood of comparability. An alternative procedure, randomized blocks assignment, or assignment by stratified blocks, involves matching (arranging) prospective participants in subgroups that contain participants that are highly comparable on key dimensions (e.g., socioeconomic status indicators) and contain the same number of participants as the number of conditions. For example, if the study requires three conditions—a standard treatment, an experimental treatment, and a waitlist condition—participants can be arranged in matching groups of three so that each trio is highly comparable on preselected features. Members in each trio are then randomly assigned to one of the three conditions, in turn increasing the probability that each condition will contain comparable participants while at the same time retaining a critical randomization procedure.

Evaluating Treatment Response Across Time

In the RCT, it is essential to evaluate participant functioning on the dependent variables

(e.g., presenting symptoms) prior to treatment initiation. Such pretreatment (or "baseline") assessments provide critical data to evaluate between-groups comparability at treatment outset, as well as within-groups treatment response. Posttreatment assessments of participants are essential to examine the comparative efficacy of treatment versus control conditions. Importantly, evidence of acute treatment efficacy (i.e., improvement immediately upon therapy completion) may not be indicative of long-term success (maintenance). At posttreatment, treatment effects may be appreciable but fail to exhibit maintenance at a follow-up assessment. Accordingly, we recommend that treatment outcome studies systematically include a follow-up assessment. Follow-up assessments (e.g., 6 months, 9 months, 18 months) are essential to demonstrations of treatment efficacy and are a signpost of methodological rigor. Maintenance is demonstrated when a treatment produces results at the follow-up assessment that are comparable to those found at posttreatment (i.e., improvements from pretreatment and an absence of detrimental change from posttreatment to follow-up).

Follow-up evaluations can help to identify differential treatment effects of considerable clinical utility. For example, two treatments may produce comparable effects at the end of treatment, but one may be more effective in the prevention of relapse (see Anderson & Lambert, 2001, for demonstration of survival analysis in clinical psychology). When two treatments show comparable response at posttreatment, yet one is associated with a higher relapse rate over time, follow-up evaluations provide critical data to support selection of one treatment over the other. For example, Brown and colleagues (1997) compared CBT and relaxation training for depression in alcoholism. When considering the average (mean) days abstinent and drinks per day as dependent variables, measured at pretreatment and at 3 and 6 months posttreatment, the authors found that although both treatments produced comparable acute gains, CBT was superior to relaxation training in maintaining the gains.

Follow-up evaluations can also be used to detect continued improvement—the benefits of some interventions may accumulate over time and possibly expand to other domains of functioning. Policymakers and researchers are increasingly interested in expanding intervention research to consider potential indirect effects on the prevention of secondary problems. In a long-term (7.4 years) follow-up of individuals treated with CBT for childhood anxiety (Kendall, Safford, Flannery-Schroeder & Webb, 2004), it was found that positive responders relative to less-positive responders had fewer problems with substance use at the long-term follow-up (see also Kendall & Kessler, 2002). In another example, participants in the Treatment for Adolescents with Depression Study (TADS) were followed for 5 years after study entry (Curry et al., 2011). TADS evaluated the relative effectiveness of fluoxetine, CBT, and their combination in the treatment of adolescents with major depressive disorder (see Treatment for Adolescents with Depression Study [TADS] Team, 2004). The Survey of Outcomes Following Treatment for Adolescent Depression (SOFTAD) was an open, 3.5-year follow-up period extending beyond the TADS 1-year follow-up period. Initial acute outcomes (measured directly after treatment) found combination treatment to be associated with significantly greater outcomes relative to fluoxetine or CBT alone, and CBT showed no incremental response over pill placebo immediately following treatment (TADS Team, 2004). However, by the 5-year follow-up, 96 percent of participants, regardless of treatment condition, experienced remission of their major depressive episode, and 88 percent recovered by 2 years (Curry et al., 2011). Importantly, gains identified at long-term follow-ups are fully attributable to the initial treatment only if one determines that participants did not seek or receive additional treatments during the follow-up interval. As an example, during the TADS 5-year follow-up interim, 42 percent of participants received psychotherapy and 44 percent received antidepressant medication (Curry et al., 2011). Appropriate statistical tests are needed to account for differences across conditions when services have been rendered during the long-term follow-up interval.

Multiple Treatment Comparisons

To evaluate the comparative (or relative) efficacy of therapeutic interventions, researchers use between-groups designs with more than one active treatment condition. Such between-groups designs offer direct comparisons of one treatment with one or more alternative active treatments. Importantly, whereas larger effect sizes can be reasonably expected in evaluations comparing an active treatment to an inactive condition, smaller differences are to be expected when distinguishing among multiple active treatments. Accordingly, sample size considerations are influenced by whether the comparison is between a treatment and a control condition or

one treatment versus another known-to-be-effective treatment (see Kazdin & Bass, 1989; see Chapter 12 in this volume for a full consideration of statistical power). Research aiming to identify reliable differences in response between two active treatments will need to evaluate a larger sample of participants than research comparing an active condition to an inactive treatment.

In a recent example utilizing multiple active treatment comparisons, Walkup and colleagues (2008) examined the efficacy of CBT, sertraline, and their combination in a placebo-controlled trial with children diagnosed with separation anxiety disorder, generalized anxiety disorder, and/or social phobia. Participants were assigned to CBT, sertraline, their combination, or a placebo pill for 12 weeks. Patients receiving the three active treatments all fared significantly better than those in the placebo group, with the combination of sertraline and CBT yielding the most favorable treatment outcomes. Specifically, analyses revealed a significant clinical response in roughly 81 percent of youth treated with a combination of CBT and sertraline, 60 percent of youth treated with CBT alone, 55 percent of youth treated with sertraline alone, and 24 percent receiving placebo alone.

As in the above-mentioned designs, it is wise to check the participant comparability across conditions on important variables (e.g., baseline functioning, prior therapy experience, socioeconomic indicators, treatment preferences/expectancies) before continuing with statistical evaluation of the intervention effects. Multiple treatment comparisons are optimal when each participant is randomly assigned to receive one and only one treatment. As previously noted, a randomized block procedure, with participants blocked on preselected variable(s) (e.g., baseline severity), can be used. Comparability across therapists who are administering the different treatments is also essential. Therapists conducting each type of treatment should be equivalent in training, experience, intervention expertise, treatment allegiance, and expectation that the intervention will be effective. To control for therapist variables, one method has each study therapist conduct each type of intervention in the study. This method is optimized when cases are randomly assigned to therapists who are equally expert and favorably disposed toward each treatment. For example, an intervention test would have reduced validity if a group of psychodynamic therapists were asked to conduct both a CBT (in which their expertise is low) and a psychodynamic therapy (in which their expertise is high). *Stratified blocking* offers a viable option to ensure that all treatments are conducted by comparable therapists. It is wise to gather data on therapist variables (e.g., expertise, experience, allegiance) and examine their relationships to participant outcomes.

For proper evaluation, intervention procedures across treatments must be equated for key variables such as (a) duration; (b) length, intensity, and frequency of contacts with participants; (c) credibility of treatment rationale; (d) treatment setting; and (e) degree of involvement of persons significant to the participant. These factors may be the basis for two alternative therapies (e.g., conjoint vs. individual marital therapy). In such cases, the nonequated feature constitutes an experimentally manipulated variable rather than a factor to control.

What is the best method of measuring change when two alternative treatments are being compared? Importantly, measures should cover the range of symptoms and functioning targeted for change, tap costs and potential negative side effects, and be unbiased with respect to the alternate interventions. Assessments should not be differentially sensitive to one treatment over another. Treatment comparisons will be misleading if measures are not equally sensitive to the types of changes that most likely result from each intervention type.

Special issues are presented in comparisons of psychological and psychopharmacological treatments (e.g., Beidel et al., 2007; Dobson et al., 2008; Marcus et al., 2007; MTA Cooperative Group, 1999; Pediatric OCD Treatment Study Team, 2004; Walkup et al, 2008). For example, when and how should placebo medications be used in comparison to or with psychological treatment? How should expectancy effects be addressed? How should differential attrition be handled statistically and/or conceptually? How should inherent differences in professional contact across psychological and pharmacological interventions be addressed? Follow-up evaluations become particularly important after acute treatment phases are discontinued. Psychological treatment effects may persist after treatment, whereas the effects of medications may not persist upon medication discontinuation. (Interested readers are referred to Hollon, 1996; Hollon & DeRubeis, 1981; Jacobson & Hollon, 1996a, 1996b, for thorough consideration of these issues.)

Procedural Considerations

We now address key RCT procedural considerations, including (a) sample selection, (b) study

setting, (c) defining the independent variable, and (d) checking the integrity of the independent variable.

Sample Selection

Selecting a sample to best represent the clinical population of interest requires careful deliberation. A *selected sample* refers to a sample of participants who may require treatment but who may otherwise only approximate clinically disordered persons. By contrast, RCTs optimize external validity when treatments are applied and evaluated with actual treatment-seeking patients. Consider a study investigating the effects of a treatment on social anxiety disorder. The researcher could use (a) a sample of patients diagnosed with social anxiety disorder via structured diagnostic interviews (*genuine clinical sample*), (b) a sample consisting of a group of individuals who self-report shyness (*analogue sample*), or (c) a sample of socially anxious persons after excluding cases with depressed mood and/or substance use (*highly select sample*). This last sample may meet full diagnostic criteria for social anxiety disorder but are nevertheless highly selected.

From a feasibility standpoint, clinical researchers may find it easier to recruit analogue samples relative to genuine clinical samples, and such samples may afford a greater ability to control various conditions and minimize threats to internal validity. At the same time, analogue and select samples compromise external validity—these individuals are not necessarily comparable to patients seen in typical clinical practice (and may not qualify as an RCT). With respect to social anxiety disorder, for instance, one could question whether social anxiety disorder in genuine clinical populations compares meaningfully to self-reported shyness (see Heiser, Turner, Beidel, & Roberson-Nay, 2009). When deciding whether to use clinical, analogue, or select samples, the researcher needs to consider how the study results will be interpreted and generalized. Regrettably, nationally representative data show that standard exclusion criteria set for clinical treatment studies exclude up to 75 percent of affected individuals in the general population who have major depression (Blanco, Olfson, Goodwin, et al., 2008).

Researchers must consider *patient diversity* when deciding which samples to study. Research supporting the efficacy of psychological treatments has historically been conducted with predominantly European-American samples, although this is rapidly changing (see Huey & Polo, 2008). Although racially and ethnically diverse samples may be similar in many ways to single-ethnicity samples, one can question the extent to which efficacy findings from predominantly European-American samples can be generalized to ethnic-minority samples (Bernal, Bonilla, & Bellido, 1995; Bernal & Scharron-Del-Rio, 2001; Hall, 2001; Olfson, Cherry, & Lewis-Fernandez, 2009; Sue, 1998). Investigations have also addressed the potential for bias in diagnoses and in the provision of mental health services to ethnic-minority patients (e.g., Flaherty & Meaer, 1980; Homma-True, Green, Lopez, & Trimble, 1993; Lopez, 1989; Snowden, 2003).

A simple rule is that the research sample should reflect the broad population to which the study results are to be generalized. To generalize to a single-ethnicity group, one must study a single-ethnicity sample. To generalize to a diverse population, one must study a diverse sample, as most RCTs strive to accomplish. Barriers to care must be reduced and outreach efforts employed to inform minorities of available services (see Sweeney, Robins, Ruberu, & Jones, 2005; Yeh, McCabe, Hough, Dupuis, & Hazen, 2003) and include them in the research. Walders and Drotar (2000) provide guidelines for recruiting and working with ethnically diverse samples.

After the fact, appropriate statistical analyses can examine potential differential outcomes (see Arnold et al., 2003; Treadwell, Flannery-Schroeder, & Kendall, 1994). Although grouping and analyzing research participants by racial or ethnic status is a common analytic approach, this approach is simplistic because it fails to address variations in each patient's degree of ethnic identity. It is often the degree to which an individual identifies with an ethnocultural group or community, and not simply his or her ethnicity itself, that may moderate response to treatment. For further consideration of this important issue, the reader is referred to Chapter 21 in this volume.

Study Setting

Some have questioned whether outcomes found at select research centers can transport to clinical practice settings, and thus the question of whether an intervention can be transported to other service settings requires independent evaluation (Southam-Gerow, Ringeisen, & Sherrill, 2006). It is not sufficient to demonstrate treatment efficacy within a narrowly defined sample in a highly selective setting. One should study, rather than assume, that a treatment found to be efficacious within a research

clinical setting will be efficacious in a clinical service setting (see Hoagwood, 2002; Silverman, Kurtines, & Hoagwood, 2004; Southam-Gerow et al., 2006; Weisz, Donenberg, Han, & Weiss, 1995; Weisz, Weiss, & Donenberg, 1992). Closing the gap between RCTs and clinical practice requires transporting effective treatments (getting "what works" into practice) and identifying additional research into those factors that may be involved in successful transportation (e.g., patient, therapist, researcher, service delivery setting; see Kendall & Southam-Gerow, 1995; Silverman et al., 2004). Methodological issues relevant to the conduct of research evaluating the transportability of treatments to "real-world" settings can be found in Chapter 5 in this volume.

Defining the Independent Variable

Proper treatment evaluation necessitates that the treatment must be adequately described and detailed in order to replicate the evaluation in another setting, or to be able to show and teach others how to conduct the treatment. Treatment manuals achieve the required description and detail of the treatment. Treatment manuals enhance internal validity and treatment integrity and allow for comparison of treatments across formats and contexts, while at the same time reducing potential confounds (e.g., differences in the amount of clinical contact, type and amount of training). Therapist manuals facilitate training and contribute meaningfully to replication (Dobson & Hamilton, 2002; Dobson & Shaw, 1988).

The merits of manual-based treatments are not universally agreed upon. Debate has ensued regarding the appropriate use of manual-based treatments versus a more variable approach typically found in clinical practice (see Addis, Cardemil, Duncan, & Miller, 2006; Addis & Krasnow, 2000; Westen, Novotny, & Thompson-Brenner, 2004). Some have argued that treatment manuals limit therapist creativity and place restrictions on the individualization that the clinicians use (see also Waltz, Addis, Koerner, & Jacobson, 1993; Wilson, 1995). Indeed, some therapy manuals may appear "cookbook-ish," and some lack attention to the clinical sensitivities needed for implementation and individualization, but our experience and data suggest that this is not the norm. An empirical evaluation from our laboratory found that the use of a manual-based treatment for child anxiety disorders (Kendall & Hedtke, 2006) did not restrict therapist flexibility (Kendall & Chu, 1999). Although it is not the goal of manual-based treatments to have clinicians perform treatment in a rigid manner, this misperception has restricted some clinicians' openness to manual-based interventions (Addis & Krasnow, 2000).

Effective use of manual-based treatments must be preceded by adequate training (Barlow, 1989). Clinical professionals cannot become proficient in the administration of therapy simply by reading a manual. Interactive training, flexible application, and ongoing clinical supervision are essential to ensure proper conduct of manual-based therapy: The goal has been referred to as "flexibility within fidelity" (Kendall & Beidas, 2007).

Several modern treatment manuals allow the therapist to attend to each patient's specific circumstances, clinical needs, concerns, and comorbid diagnoses without deviating from the core treatment strategies detailed in the manual. The goal is to include provisions for standardized implementation of therapy while using a personalized case formulation (e.g., see Suveg, Comer, Furr, & Kendall, 2006). Importantly, use of manual-based treatments does not eliminate the potential for differential therapist effects. Researchers examine therapist variables within the context of manual-based treatments (e.g., therapeutic relationship-building behaviors, flexibility, warmth) that may relate to treatment outcome (Creed & Kendall, 2005; Karver et al., 2008; Shirk et al., 2008; see also Chapter 9 in this volume for a full consideration of designing, conducting, and evaluating therapy process research).

Checking the Integrity of the Independent Variable

Careful checking of the manipulated variable is required in any rigorous experimental research. In the RCT, the manipulated variable is typically treatment or a key characteristic of treatment. By experimental design, all participants are not treated the same. However, just because the study has been so designed does not guarantee that the independent variable (treatment) has been implemented as intended. In the course of a study—whether due to insufficient therapist training, therapist variables, lack of manual specification, inadequate therapist monitoring, participant demand characteristics, or simple error variance—the treatment that was assigned may not in fact be the treatment that was provided (see also Perepletchikova & Kazdin, 2005).

To help ensure that the treatments are indeed implemented as intended, it is wise to require that a treatment plan be followed, that therapists are

carefully trained, and that sufficient supervision is available throughout. The researcher is wise to conduct an independent check on the manipulation. For example, treatment sessions are recorded so that an independent rater can listen to and/or watch the recordings and provide quantifiable judgments regarding key characteristics of the treatment. Such a manipulation check provides the necessary assurance that the described treatment was indeed provided as intended. Digital audio and video recordings are inexpensive, can be used for subsequent training, and can be analyzed to answer key research questions. Therapy session recordings evaluated in RCTs not only provide a check on the treatment within each separate study but also allow for a check on the comparability of treatments provided across studies. That is, the therapy provided as CBT in one researcher's RCT could be checked to assess its comparability to other teams' CBT.

A recently completed clinical trial from our research program comparing two active-treatment conditions for childhood anxiety disorders against an active attention control condition (Kendall et al., 2008) illustrates a procedural plan for integrity checks. First, we developed a checklist of the strategies and content called for in each session by the respective treatment manuals. A panel of expert clinicians served as independent raters who used the checklists to rate randomly selected video segments from randomly selected cases. The panel of raters was trained on nonstudy cases until they reached an interrater reliability of Cohen's $\kappa \geq .85$. After ensuring reliability, the panel used the checklists to assess whether the appropriate content was covered for randomly selected segments that were representative of all sessions, conditions, and therapists. For each coded session, we computed an integrity ratio corresponding to the number of checklist items covered by the therapist divided by the total number of items that should have been included. Integrity check results indicated that across the conditions, 85 to 92 percent of intended content was in fact covered.

It is also wise for the RCT researcher to evaluate the *quality* of treatment provided. A therapist may strictly adhere to a treatment manual and yet fail to administer the treatment in an otherwise competent manner, or he or she may administer therapy while significantly deviating from the manual. In both cases, the operational definition of the independent variable (i.e., the treatment manual) has been violated, treatment integrity impaired, and replication rendered impossible (Dobson & Shaw, 1988). When a treatment fails to demonstrate expected gains, one can examine the adequacy with which the treatment was implemented (see Hollon, Garber, & Shelton, 2005). It is also of interest to investigate potential variations in treatment outcome that may be associated with differences in the *quality* of the treatment provided (Garfield, 1998; Kendall & Hollon, 1983). Expert judges are needed to make determinations of differential quality prior to the examination of differential outcomes for high- versus low-quality therapy implementation (see Waltz et al., 1993). McLeod and colleagues (in press) provide a description of procedural issues in the conduct of quality assurance and treatment integrity checks.

Measurement Considerations

Assessing the Dependent Variable(s)

No single measure can serve as the sole indicator of participants' treatment-related gains. Rather, a variety of methods, measures, data sources, and sampling domains (e.g., symptoms, distress, functional impairment, quality of life) are used to assess outcomes. A rigorous treatment RCT will consider using assessments of participant self-report; participant test/task performance; therapist judgments and ratings; archival or documentary records (e.g., health care visits and costs, work and school records); observations by trained, unbiased, blinded observers; rating by significant people in the participant's life; and independent judgments by professionals. Outcomes are more compelling when observed by independent (blind) evaluators than when based solely on the therapist's opinion or the participant's self-reports.

Collecting data on variables of interest from multiple reporters (e.g., treatment participant, family members, peers) can be particularly important when assessing children and adolescents. Such a *multi-informant strategy* is critical as features of cognitive development may compromise youth self-reports, and children may simply offer what they believe to be the desired responses. And so in RCTs with youth, collecting additional data from important adults in children's lives who observe them across different settings (e.g., parents, teachers) is essential. Importantly, however, because emotions and mood are partially internal phenomena, some symptoms may be less known to parents and teachers, and some observable symptoms may occur in situations outside the home or school. Accordingly, an inherent dilemma with a multi-informant assessment strategy is that discrepancies among informants are to be expected (Comer & Kendall, 2004). Research shows low to moderate concordance rates

among informants in the assessment of youth (De Los Reyes & Kazdin, 2005), with particularly low agreement among child internalizing symptoms (Comer & Kendall, 2004).

A *multimodal strategy* relies on multiple inquiries to evaluate an underlying construct of interest. For example, assessing family functioning may include family members completing self-report forms on their perceptions of relationships in the family, as well as conducting structured behavioral observations of family members interacting to be coded by independent raters. Statistical packages can integrate data obtained from multimodal assessment strategies (see Chapter 16 in this volume). The increasing availability of handheld communication devices and personal digital assistants allows researchers to incorporate experience sampling methodology (ESM), in which people report on their emotions and behavior in real-world situations (*in situ*). ESM data provide naturalistic information on patterns in day-to-day functioning (see Chapter 11 in this volume).

In a well-designed RCT, multiple targets are assessed to determine treatment evaluation. For example, one can measure the presence of a diagnosis, overall well-being, interpersonal skills, self-reported mood, family functioning, occupational impairment, and health-related quality of life. No one target captures all, and using multiple targets facilitates an examination of therapeutic changes when changes occur, and the absence of change when interventions are less beneficial. However, inherent in a multiple-domain assessment strategy is the fact that it is rare that a treatment produces uniform effects across assessed domains. Suppose a treatment, relative to a control condition, improves participants' severity of anxiety, but not their overall quality of life. In an RCT designed to evaluate improved anxiety symptoms and quality of life, should the treatment be deemed efficacious if only one of two measures showed gains? The Range of Possible Changes model (De Los Reyes & Kazdin, 2006) calls for a multidimensional conceptualization of intervention change. In this spirit, we recommend that RCT researchers be explicit about the domains of functioning expected to change and the relative magnitude of such expected changes. We also caution consumers of the treatment outcome literature against simplistic dichotomous appraisals of treatments as efficacious or not.

Data Analysis

Data analysis is an active process through which we extract useful information from the data we have collected in ways that allow us to make statistical inferences about the larger population that a given sample was selected to represent. Data do not "speak" for themselves. Although a comprehensive statistical discussion about RCT data analysis is beyond the present scope (the reader is referred to Jaccard & Guilamo-Ramos, 2002a, 2002b; Kraemer & Kupfer, 2006; Kraemer, Wilson, Fairburn, & Agras, 2002; and Chapters 14 and 16 in this volume) in this section, we discuss three areas that merit consideration in the context of RCT data analysis: (a) addressing missing data and attrition, (b) assessing clinical significance, and (c) evaluating mechanisms of change (i.e., mediators and moderators).

Addressing Missing Data and Attrition

Not every participant assigned to treatment actually completes participation in an RCT. A loss of research participants (*attrition*) may occur just after randomization, during treatment, prior to posttreatment evaluation, or during the follow-up interval. Increasingly, researchers are evaluating predictors and correlates of attrition to elucidate the nature of treatment dropout, to understand treatment tolerability, and to enhance the sustainability of mental health services in the community (Kendall & Sugarman, 1997; Reis & Brown, 2006; Vanable, Carey, Carey, & Maisto., 2002). However, from a research methods standpoint, attrition can be problematic for data analysis, such as when there are large numbers of noncompleters or when attrition varies across conditions (Leon et al., 2006; Molenberghs et al., 2004).

Regardless of how diligently researchers work to prevent attrition, data will likely be lost. Although attrition rates vary across RCTs and treated clinical populations, Mason (1999) estimated that most researchers can expect roughly 20 percent of their sample to withdraw or be removed from a study prior to completion. To address this matter, researchers can conduct and report two sets of analyses: (a) analyses of outcomes for treatment completers and (b) analyses of outcomes for all participants who were included at the time of randomization (i.e., the *intent-to-treat sample*). Treatment-completer analyses involve the evaluation of only those who actually completed treatment and examine what the effects of treatment are when someone completes a full treatment course. Treatment refusers, treatment dropouts, and participants who fail to adhere to treatment schedules are not included in such analyses. Reports of such treatment outcomes may be somewhat elevated because they represent

the results for only those who adhered to and completed the treatment. A more conservative approach to addressing missing data, intent-to-treat analysis, entails the evaluation of outcomes for all participants involved at the point of randomization. As proponents of intent-to-treatment analyses we say, "once randomized, always analyzed."

Careful consideration is required when selecting an appropriate analytic method to handle missing endpoint data because different methods can produce different outcomes (see Chapter 19 in this volume). Researchers address missing endpoint data via one of several ways: (a) *last observation carried forward* (LOCF), (b) substituting pretreatment scores for posttreatment scores, (c) multiple imputation methods, and (d) mixed-effects models. LOCF analysis assumes that participants who drop out remain constant on the outcome variable from their last assessed point through the posttreatment evaluation. For example, if a participant drops out at week 6, the data from the week 5 assessment would be substituted for his or her missing posttreatment assessment data. The LOCF approach can be problematic however, as the last data collected may not be representative of the dropout participant's ultimate progress or lack of progress at posttreatment, given that participants may change after dropping out of treatment. The use of pretreatment data as posttreatment data (a conservative and not recommended method) simply inserts pretreatment scores for cases of attrition as posttreatment scores, assuming that participants who drop out make no change from their initial baseline state. Critics of pretreatment substitution and LOCF argue that these crude methods introduce systematic bias and fail to take into account the uncertainty of posttreatment functioning (see Leon et al., 2006). More current missing data imputation methods are grounded in statistical theory and incorporate the uncertainty regarding the true value of the missing data. *Multiple imputation methods* impute a range of values for the missing data, incorporating the uncertainty of the true values of missing data and generating a number of nonidentical datasets (Little & Rubin, 2002). After the researcher conducts analyses on the nonidentical datasets, the results are pooled and the resulting variability addresses the uncertainty of the true value of the missing data.

Mixed-effects modeling, which relies on linear and/or logistic regression to address missing data in the context of random (e.g., participant) and fixed effects (e.g., treatment, age, sex) (see Hedeker & Gibbons, 1994, 1997; Laird & Ware, 1982), can be used (see Neuner et al., 2008, for an example). Mixed-effects modeling may be particularly useful in addressing missing data if numerous assessments are collected across a treatment trial (e.g., weekly symptom reports are collected).

Despite sophisticated data analytic approaches to accounting for missing data, we recommend that researchers attempt to contact noncompleting participants and re-evaluate them at the time when the treatment would have ended. This method accounts for the passage of time, as both dropouts and treatment completers are evaluated over time periods of the same duration, and minimizes any potential error introduced by statistical imputation and modeling approaches to missing data. If this method is used, however, it is important to determine whether dropouts sought and/or received alternative treatments in the interim.

Assessing the Persuasiveness of Therapeutic Outcomes

Data produced by RCTs are submitted to statistical tests of significance. Mean scores for participants in each condition are compared, within-group and between-group variability is considered, and the analysis produces a numerical figure, which is then checked against critical values. *Statistical* significance is achieved when the magnitude of the mean difference is beyond that which could have resulted by chance alone (conventionally defined as $p < .05$). Tests of statistical significance are essential as they inform us that the degree of change was likely not due to chance.

Importantly, statistical tests alone do not provide evidence of *clinical significance*. Sole reliance on statistical significance can lead to perceiving treatment gains as potent when in fact they may be clinically insignificant. For example, imagine that the results of a treatment outcome study demonstrate that mean Beck Depression Inventory (BDI) scores are significantly lower at posttreatment than pretreatment. An examination of the means, however, reveals only a small but reliable shift from a mean of 29 to a mean of 26. With larger sample sizes, this difference may well achieve statistical significance at the conventional $p < .05$ level (i.e., over 95 percent chance that the finding is not due to chance alone), yet perhaps be of limited practical significance. Both before and after treatment, the scores are within the range considered indicative of clinical levels of depressive distress (Kendall, Hollon, Beck, Hammen, & Ingram, 1987), and such a small magnitude of change may have little effect on a person's

life impairment (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). Conversely, statistically meager results may disguise meaningful changes in participant functioning. As Kazdin (1999) put it, sometimes a little can mean a lot, and vice versa.

Clinical significance refers to the persuasiveness or meaningfulness of the magnitude of change (Kendall, 1999). Whereas statistical significance tests address the question, "Were there treatment-related changes?" tests of clinical significance address the question, "Were treatment-related changes meaningful and convincing?" Specifically, this can be made operational as changes on a measure of the presenting problem (e.g., anxiety symptoms) that result in the participants being returned to within normal limits on that same measure. Several approaches for measuring clinically significant change have been developed, two of which are *normative sample comparison* and *reliable change index*.

Normative comparisons (Kendall & Grove, 1988; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999) are conducted in several steps. First, the researcher selects a normative group for posttreatment comparison. Given that several well-established measures provide normative data (e.g., the BDI, the Child Behavior Checklist), investigators may choose to rely on these preexisting normative samples. However, when normative data do not exist, or when the treatment sample is qualitatively different on key factors (e.g., socioeconomic status indicators, age), it may be necessary to collect one's own normative data. In a typical RCT, when using statistical tests to compare groups, the investigator assumes equivalency across groups (null hypothesis) and aims to find that they are not (alternate hypothesis). However, when the goal is to show that treated individuals are equivalent to "normal" individuals on some factor (i.e., are indistinguishable from normative comparisons), traditional hypothesis-testing methods are inadequate. One uses an equivalency testing method to circumvent this problem (Kendall, Marrs-Garcia, et al., 1999) that examines whether the difference between the treatment and normative groups is within some predetermined range. When used in conjunction with traditional hypothesis testing, this approach allows conclusions to be drawn about the equivalency of groups (see, e.g., Jarrett, Vittengl, Doyle, & Clark, 2007; Pelham et al., 2000; Westbrook & Kirk, 2007, for examples of normative comparisons), thus testing that posttreatment data are within a normative range on the measure of interest. For example, Weisz and colleagues (1997) utilized normative comparisons in a trial in which elementary school children with mild to moderate symptoms of depression were randomly assigned either to a Primary and Secondary Control Enhancement Training (PASCET) program or to a no-treatment control group. Normative comparisons were used to determine whether participants' scores on two depression measures, the Children's Depression Inventory and the Revised Children's Depression Rating Scale, fell within one standard deviation above elementary school norm groups at pretreatment, posttreatment, and 9-month follow-up time points. Utilizing normative comparisons allowed the authors to conclude that children who had received the treatment intervention were more likely to fall within the normal range on depression measures than children in the no-treatment control condition.

The Reliable Change Index (RCI; Jacobson, Follette, & Revenstorf, 1984; Jacobson & Traux, 1991) is another popular method to examine clinically significant change. The RCI entails calculating the number of participants moving from a dysfunctional to a normative range. Specifically, the research calculates a difference score (posttreatment minus pretreatment) divided by the standard error of measurement (calculated based on the reliability of the measure). The RCI is influenced by the magnitude of change and the reliability of the measure. The RCI has been used in RCT research, although its originators point out that it has at times been misapplied (Jacobson, Roberts, Berns, & McGlinchey, 1999). When used in conjunction with reliable measures and appropriate cutoff scores, it can be a valuable tool for assessing clinical significance.

Evaluating Change Mechanisms

The RCT researcher is often interested in identifying (a) the conditions that dictate when a treatment is more or less effective and (b) the processes through which a treatment produces change. Addressing such issues necessitates the specification of *moderator* and *mediator* variables (Baron & Kenny, 1986; Holmbeck, 1997; Kraemer et al., 2002). A moderator is a variable that delineates the conditions under which a given treatment is related to an outcome. Conceptually, moderators identify *on whom* and *under what circumstances* treatments have different effects (Kraemer et al., 2002). A moderator is functionally a variable that influences either the strength or direction of a relationship between an independent variable (treatment) and a dependent variable (outcome). For example, if in an RCT the experimental treatment was found

to be more effective with men than with women, but this gender effect was not found in response to the control treatment, then gender would be considered a moderator of the association between treatment and outcome. Treatment moderators help clarify for consumers of the treatment outcome literature which patients might be most responsive to which treatments, and for which patients alternative treatments might be sought. Importantly, when a variable broadly predicts outcome across all treatment conditions in an RCT, conceptually that variable is simply a *predictor*, and not a moderator (see Kraemer et al., 2002).

On the other hand, a mediator is a variable that serves to explain the process by which a treatment affects an outcome. Conceptually, mediators identify *how* and *why* treatments take effect (Kraemer et al., 2002). The mediator effect reveals the mechanism through which the independent variable (e.g., treatment) is related to outcome (e.g., treatment-related changes). Accordingly, mediational models are inherently causal models, and in the context of an RCT, significant meditational pathways inform us about causal relationships. If an effective treatment for child externalizing problems was found to have an impact on parenting behavior, which in turn was found to have a significant influence on child externalizing behavior, then parent behavior would be considered to mediate the treatment-to-outcome relationship (provided certain statistical criteria were met; see Holmbeck, 1997). Specific statistical methods used to evaluate the presence of treatment moderation and mediation can be found elsewhere (see Chapter 15 in this volume).

Reporting the Results

Communicating study findings to the scientific community constitutes the final stage of conducting an RCT. A quality report will present outcomes in the context of previous related work (e.g., discussing how the findings build on and support previous work; discussing the ways in which findings are discrepant from previous work and why this may be the case), as well as consider shortcomings and limitations that can direct future empirical efforts and theory in the area. To prepare a well-constructed report, the researcher must provide all of the relevant information for the reader to critically appraise, interpret, and/or replicate study findings. It has been suggested that there have been inadequacies in the reporting of RCTs (see Westen et al., 2004). Inadequacies in the reporting of RCTs can result in bias in estimating treatment effectiveness (Moher, Schulz, & Altman, 2001). An international group of epidemiologists, statisticians, and journal editors developed a set of consolidated standards for reporting trials (i.e., CONSORT; see Begg et al., 1996) in order to maximize transparency in RCT reporting. CONSORT guidelines consist of a 22-item checklist of study features that can bias estimates of treatment effects, or that are critical to judging the reliability or relevance of RCT findings, and consequently should be included in a comprehensive research report. A quality report will address each of these 22 items. Importantly, participant flow should be characterized at each research stage. The researcher reports the specific numbers of participants who were randomly assigned to each treatment condition, who received treatments as assigned, who participated in post-treatment evaluations, and who participated in follow-up evaluations. It has become standard practice for scientific journals to require a CONSORT flow diagram. See Figure 4.1 for an example of a flow diagram used in reporting to depict participant flow at each stage of an RCT.

Next, the researcher must decide where to submit the report. We recommend that researchers consider submitting RCT findings to peer-reviewed journals only. Publishing RCT outcomes in a refereed journal (i.e., one that employs the peer-review process) signals that the work has been accepted and approved for publication by a panel of impartial and qualified reviewers (i.e., independent researchers knowledgeable in the area but not involved with the RCT). Consumers should be highly cautious of RCTs published in journals that do not place manuscript submissions through a rigorous peer-review process. Although the peer-review process slows down the speed with which one is able to communicate RCT results, much to the chagrin of the excited researcher who just completed an investigation, it is nonetheless one of the indispensable safeguards that we have to ensure that our collective knowledge base is drawn from studies meeting acceptable standards. Typically, the review process is "blind," meaning that the authors of the article do not know the identities of the peer reviewers who are considering their manuscript. Many journals now employ a double-blind peer-review process in which the identities of study authors are also not known to the peer reviewers.

Extensions and Variations of the RCT

Thus far, we have addressed considerations related to the design and implementation of the standard RCT. We now turn our attention to important extensions and variations of the RCT. These

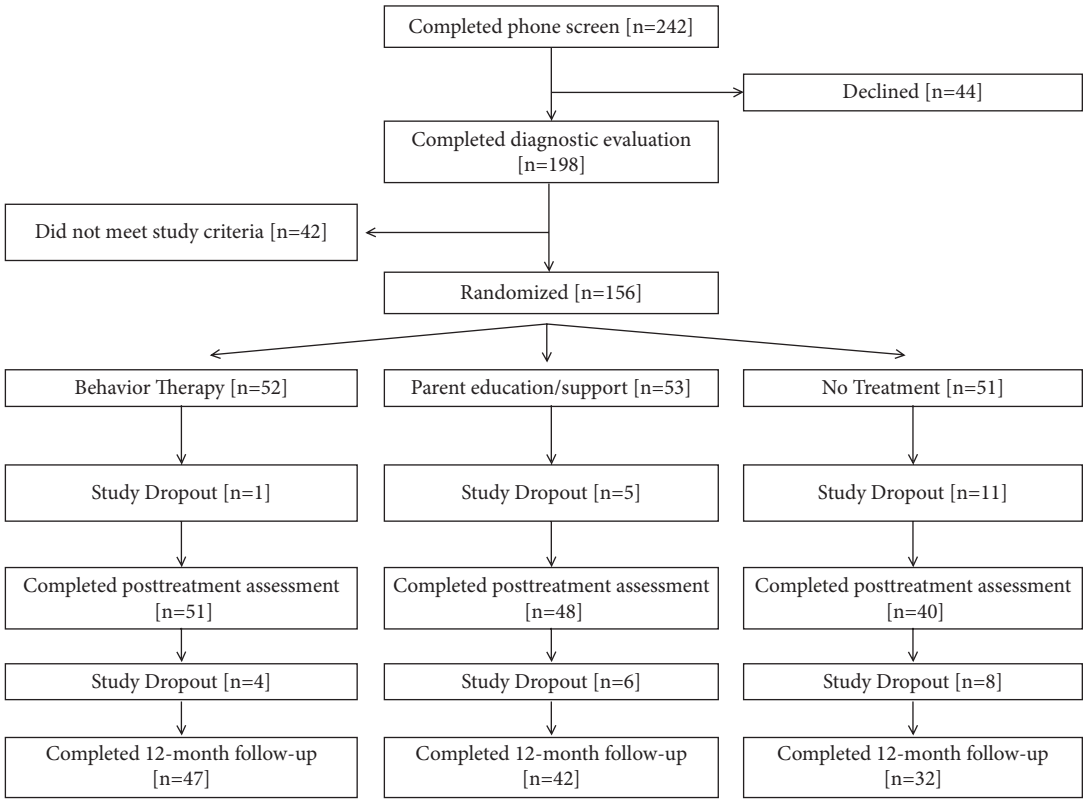


Figure 4.1 Example of flow diagram used in reporting to depict participant flow at each study stage.

treatment study designs—which include equivalency designs and sequenced treatment designs—address key questions that cannot be adequately addressed with the traditional RCT. We discuss each of these designs in turn and note some of their strengths and limitations.

Equivalency Designs

As varied therapeutic treatment interventions are becoming readily available, research is needed to determine their relative efficacy. Sometimes, the researcher is not interested in evaluating the superiority of one treatment over another, but rather that a treatment produces comparable results to another treatment that differs in key ways. For example, a researcher may be interested in determining whether an individual treatment protocol can yield comparable results when administered in a group format. The researcher may not hold a hypothesis that the group format would produce superior outcomes, but if it could be demonstrated that the two treatments produce equivalent outcomes, the group treatment may nonetheless be preferred due to the efficiency of treating multiple patients in the same amount of time. In another example, a researcher may be interested in comparing a cross-diagnostic treatment (e.g., one that flexibly addresses any of the common child anxiety disorders—separation anxiety disorder, social anxiety disorder, or generalized anxiety disorders) relative to single-disorder treatment protocols for those specific disorders. The researcher may not hold a hypothesis that the cross-diagnostic protocol produces superior outcomes over single-disorder treatment protocols, but if it could be demonstrated that it produces equivalent outcomes, parsimony would suggest that the cross-diagnostic protocol would be the most efficient to broadly disseminate.

In equivalency research designs, significance tests are utilized to determine the equivalence of treatment outcomes observed across multiple active treatments. While a standard significance test would be used in a comparative trial, such a test could not conclude equivalency between treatments because a nonsignificant difference does not necessarily signify equivalence (Altman & Bland, 1995). In an equivalency design, a confidence interval is established to define a range of points within which

treatments may be deemed essentially equivalent (Jones, Jarvis, Lewis, & Ebbutt, 1996). To minimize bias, this confidence interval must be determined prior to data collection.

Barlow and colleagues, for example, are currently testing the efficacy of a transdiagnostic treatment (Unified Protocol for Emotional Disorders; Barlow, Farchione, Fairholme, Ellard, Boisseau, et al., 2010) for anxiety disorders. The proposed analyses include a rigorous comparison of the Unified Protocol (UP) against single-diagnosis psychological treatment protocols (SDPs). Statistical equivalence will be used to test the hypothesis that the UP is statistically equivalent to SDPs. An *a priori* confidence interval around change in the clinical severity rating (CSR) will be utilized to evaluate statistical equivalence among treatments. The potential finding that the UP is indeed equivalent to SDPs in the treatment of anxiety disorders, regardless of specific diagnosis, would have important implications for treatment dissemination and transportability.

A variation of the equivalency research design is the *benchmarking design,* which involves a quantitative comparison between treatment outcomes collected in a current study and results from similar treatment outcome studies. Demonstrating equivalence in such a study design allows the researcher to determine whether results from a current treatment evaluation are equivalent to findings reported elsewhere in the literature. Results of a trial are evaluated, or benchmarked, against the findings from other comparable trials. Weersing and Weisz (2002) used a benchmarking design to assess differences in the effectiveness of community psychotherapy for depressed youth versus evidence-based CBT provided in RCTs. The authors aggregated data from all available clinical trials evaluating the effects of best-practice treatment, determined the pooled effect sizes associated with depressed youth treated in these clinical trials, and benchmarked these data with outcomes of depressed youth treated in community mental health clinics. They found that outcomes of youth treated in community care settings were more similar to youth in control conditions than to youth treated with CBT.

Benchmarking equivalency designs allow for meaningful comparison groups with which to gauge the progress of treated participants in a clinical trial. The comparison data are typically readily available, given that they may include samples that have been used to obtain normative data for specific measures, or research participants whose outcome data are included in reported results in published studies. In addition, as noted earlier, equivalency tests can be conducted to determine the clinical significance of treatment outcomes—that is, the extent to which posttreatment functioning identified in a treated group is comparable to functioning among normative comparisons (Kendall, Marrs-Garcia, et al., 1999).

Sequenced Treatment Designs

When interventions are applied, a treated participant's symptoms may improve (treatment response), may get worse (deterioration), may neither improve nor deteriorate (treatment nonresponse), or may improve somewhat but not to a satisfactory extent (partial response). In clinical practice, over the course of treatment important clinical decisions must be made regarding when to escalate treatment, augment treatment with another intervention, or switch to another supported intervention. The standard RCT design does not provide sufficient data with which to inform the optimal sequence of treatment for cases of nonresponse, partial response, or deterioration.

When the aim of a research study is to determine the most effective sequence of treatments for an identified patient population, a *sequenced treatment design* may be utilized. This design involves the assignment of study participants to a particular sequence of treatment and control/comparison conditions. The order in which conditions are assigned may be random, as in a *randomized sequence design*. In other sequenced treatment designs, factors such as participant characteristics, individual treatment outcomes, or participant preferences may influence the sequence of administered treatments. These variations on sequenced treatment designs—*prescriptive, adaptive,* and *preferential treatment designs*, respectively—are outlined in further detail below.

The *prescriptive treatment design* recognizes that individual patient characteristics play a key role in treatment outcomes and assigns treatment condition based on these patient characteristics. The basis of this treatment design aims to improve upon nomothetic data models by incorporating idiographic data to treatment assignments (see Barlow & Nock, 2009). Study participants who are matched to treatment conditions based on individual characteristics (e.g., psychiatric comorbidity, levels of distress and impairment, readiness to change, etc.) may experience greater gains than those who are not matched to interventions based on patient characteristics (Beutler & Harwood, 2000). In a prescriptive

treatment design, the clinical researcher studies the effectiveness of a treatment decision-making algorithm as opposed to a set treatment protocol. Participants do not have an equal chance of receiving study treatments, as is the case in the standard RCT. Instead, what remains consistent across participants is the application of the same decision-making algorithm, which can lead to a variety of sequenced treatment courses.

Although a prescriptive treatment design may enhance clinical generalizability—as practitioners will typically incorporate patient characteristics into treatment planning—this design introduces serious threats to internal validity. In a variation, the *randomized prescriptive treatment design* randomizes participants to either a blind randomization algorithm or an experimental treatment algorithm. For example, algorithm A may randomly assign participants to one of three treatment conditions (i.e., the blind randomization algorithm), and algorithm B may match participants to each of the three treatment conditions based on baseline data hypothesized to inform the optimal treatment assignment (i.e., the experimental treatment algorithm). Here, the researcher is interested in which algorithm condition is superior, rather than what is the absolute effect of a specific treatment protocol.

One of the primary goals of prescriptive treatment research designs is to examine the effectiveness of treatments tailored to individuals for whom those treatments are thought to work best. Key patient dimensions that have been found in nomothetic evaluations to be effective mediators or moderators of treatment outcome can lay the groundwork for decision rules to assign participants to particular interventions, or alternatively, can lead to the administration or omission of a specific module within a larger intervention. Prescriptive treatment designs offer opportunities to better develop and tailor efficacious treatments to patients with varied characteristics.

In the *adaptive treatment design,* a participant's course of treatment is determined by his or her clinical response across the trial. In the traditional RCT, a comparison is typically made between an innovative treatment and some sort of placebo or accepted standard. Some argue that a more clinically relevant design involves a comparison between an innovative treatment and an adaptive strategy in which a participant's treatment condition is switched based on treatment outcome to date (Dawson & Lavori, 2004). With the adaptive treatment study design, clinical researchers can also switch participants from one experimental group to another if a particular intervention is lacking in effectiveness for an individual patient (see Chapter 5 in this volume). After a participant reaches a predetermined deterioration threshold, or if he or she fails to meet a response threshold before a given point during a trial, the participant may be switched from the innovative treatment to the accepted standard, or vice versa. In this way, the adaptive treatment option allows the clinical researcher to determine the relative efficacy of the innovative treatment if the adaptive strategy produces significantly better outcomes than the standard treatment (Dawson & Lavori, 2004).

Illustrating the utility of the adaptive treatment design, the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study assessed the effectiveness of depression treatments in patients with major depressive disorder (Rush et al., 2004). Participants advanced through four levels of treatment and were assigned a particular course of treatment depending on their response to treatment up until that point. In level 1, all participants were given citalopram for 12 to 14 weeks. Those who became symptom-free after this period could continue on citalopram for a 12-month follow-up period, while those who did not become symptom-free moved on to level 2. Levels 2 and 3 allowed participants to choose another medication or cognitive therapy (switch) or augment their current medication with another medication or cognitive therapy (add-on). In level 4, participants who were not yet symptom-free were taken off their medications and randomly assigned to either a monoamine oxidase inhibitor (MAOI) or the combination of venlafaxine extended release with mirtazapine. At each stage of the study, participants were assigned to treatments based on their previous treatment responses. Roughly half of the participants became symptom-free after two treatment levels, and roughly 70 percent of study completers became symptom-free over the course of all four treatment levels.

The Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD), a national, longitudinal public health initiative, was implemented in an effort to gauge the effectiveness of treatments for adults with bipolar disorder (Sachs et al., 2003). All participants were initially invited to enter Standard Care Pathways (SCPs), which involved clinical care delivered by a STEP-BD clinician. At any point during their participation in SCP treatment, participants could become eligible for one of the Randomized Care Pathways (RCPs) for acute depression, refractory depression, or relapse

prevention. Upon meeting nonresponse criteria (i.e., failure to respond to treatment within the first 12 weeks, or failure to respond to two or more antidepressants in the current depressive episode), participants were randomly assigned to one of the RCP treatment arms. After participating in one of these treatment arms, participants could return to SCP or opt to participate in another RCP. Some of the treatment arms also allowed the treating clinician to exclude a participant from an RCP based on his or her particular presentation. The treatment was therefore adaptive in nature, allowing for flexibility and some element of decision making to occur within the trial. Importantly, although such flexibility may enhance generalizability to clinical settings and when used appropriately can guide clinical practice, this flexibility introduces serious threats to internal validity. Accordingly, this design does not allow inferences to be made about the absolute benefit of various interventions.

The *preferential treatment design* allows study participants to choose the treatment condition(s) to which they are assigned. This approach considers patient preferences, which emulates the process that typically occurs in clinical practice. Taking into account patient preferences in a treatment study can result in a better understanding of which individuals will fare best when administered specific interventions under circumstances that incorporate their preferences in determining treatment selection. Proponents often argue that assigning treatments based on patient preference may increase other factors known to positively affect treatment outcomes, including patient motivation, attitudes toward treatment, and expectations of treatment success.

Lin and colleagues (2005) utilized a preferential treatment design to explore the effects of matching patient preferences and interventions in a population of adults with major depression. Participants were offered antidepressant medication and/or counseling based on patient preference, where appropriate. Participants who were matched to their treatment preference exhibited more positive treatment outcomes at 3- and 9-month follow-up evaluations than participants who were not matched to their preferred treatment condition.

Importantly, outcomes identified in preferential treatment designs are intertwined with the confound of patient preferences. Accordingly, clinical researchers are wise to use preferential treatment designs only after treatment efficacy has first been established for the various treatment arms in a randomized design.

In a *multiple-groups crossover design*, participants are randomly assigned to receive a sequence of at least two treatments, one of which may be a control condition. In this design, participants act as their own controls, as at some point during the trial, they receive each of the experimental and control/comparison conditions. Because each participant is his or her own control, the risk of having comparison groups that are dissimilar on variables such as demographic characteristics, severity of presenting symptoms, and comorbidities is eliminated. Precautions should be taken to ensure that the effects of one treatment intervention have receded before starting participants on the next treatment intervention.

Illustrations of multiple-groups crossover designs can often be found in clinical trials testing the efficacy of various medications. Hood and colleagues (2010) utilized a double-blind crossover design in a study with untreated and selective serotonin reuptake inhibitor (SSRI)-remitted patients with social anxiety disorder. Participants were administered a single dose of either pramipexole (a dopamine agonist) or sulpiride (a dopamine antagonist). One week later, participants received a single dose of the medication they had not received the previous week. Following each medication administration, participants were asked to rate their anxiety and mood, and they were invited to engage in anxiety-provoking tasks. The authors concluded that untreated participants experienced significant increases in anxiety symptoms following anxiety-provoking tasks after both medications. In contrast, SSRI-remitted participants experienced elevated anxiety under the effects of sulpiride and decreased anxiety levels under the effects of pramipexole.

Multiple-groups crossover designs are best suited for the evaluation of interventions that would not expectedly retain effects once they are removed, as is the case in the evaluation of a therapeutic medication with a very short half-life. These designs are more difficult to implement in the evaluation of psychosocial interventions, which often produce effects that are somewhat irreversible (e.g., the learning of a skill, or the acquisition of important knowledge). How can the clinical researcher evaluate separate treatment phases when it is not possible to completely remove the intervention? In such situations, crossover designs are misguided.

Proponents of sequential designs argue that designs that are informed by patient characteristics, outcomes, and preferences provide patients with uniquely individualized care within a clinical trial. The argument suggests that an appropriate match

between patient characteristics and treatment type will optimize success in producing significant treatment effects and lead to a heightened understanding of interventions that are best suited to a variety of patients and circumstances in clinical practice (Luborsky et al., 2002). In this way, systematic evaluation is extended to the very decision-making algorithms that occur in real-world clinical practice, an important element not afforded by the standard RCT. However, whereas these approaches increase clinical relevance and may enhance the ability to generalize findings from research to clinical practice, they also decrease scientific rigor by eliminating the uniformity of randomization to experimental conditions.

Conclusion

The RCT offers the most rigorous method of examining the causal impact of therapeutic interventions. After reviewing the essential considerations relevant for matters of design, procedure, measurement, and data analysis, one recognizes that no one single clinical trial, even with optimal design and procedures, can alone answer the relevant questions about the efficacy and effectiveness of therapy. Rather, a series and collection of studies, with varying designs and approaches, is necessary. The extensions and variations of the RCT addressed in this chapter address important features relevant to clinical practice that are not informed by the standard RCT; however, each modification to the standard RCT decreases the maximized internal validity achieved in the standard RCT. Criteria for determining evidence-based practice have been proposed (American Psychological Association, 2005, Chambless & Hollon, 1998), and the quest to identify such treatments continues. The goal is for the research to be rigorous, with the end goal being to optimize clinical decision making and practice for those affected by emotional and behavioral disorders and problems.

Controlled trials play a vital role in facilitating a dialogue between academic clinical psychology and the public and private sector (e.g., insurance payers, Department of Health and Human Services, policymakers). The results of controlled clinical evaluations are increasingly being examined by both professional associations and managed care organizations with the intent of formulating clinical practice guidelines for cost-effective care that provides optimized service to those in need. In the absence of compelling data, there is the risk that psychological practice will be co-opted and exploited in the service of only profitability and/or cost containment. Clinical trials must retain scientific rigor to enhance the ability of practitioners to deliver effective treatment procedures to individuals in need.

References

Addis, M., Cardemil, E. V., Duncan, B., & Miller, S. (2006). Does manualization improve therapy outcomes? In J. C. Norcross, L. E. Beutler, & R. F. Levant (Eds.), *Evidence-based practices in mental health* (pp. 131–160). Washington, DC: American Psychological Association.

Addis, M., & Krasnow, A. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology*, *68*, 331–339.

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, *311* (7003), 485.

American Psychological Association. (2005). *Policy statement on evidence-based practice in psychology.* Retrieved August 27, 2011, from http://www.apa.org/practice/resources/evidence/evidence-based-statement.pdf.

Anderson, E. M., & Lambert, M. J. (2001). A survival analysis of clinical significant change in outpatient psychotherapy. *Journal of Clinical Psychology*, *57*, 875–888.

Arnold, L. E., Elliott, M., Sachs, L., Bird, H., Kraemer, H. C., Wells, K. C., et al. (2003). Effects of ethnicity on treatment attendance, stimulant response/dose, and 14-month outcome in ADHD. *Journal of Consulting and Clinical Psychology*, *71*, 713–727.

Barlow, D. H. (1989). Treatment outcome evaluation methodology with anxiety disorders: Strengths and key issues. *Advances in Behavior Research and Therapy*, *11*, 121–132.

Barlow, D. H., Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Allen, L. B., & Ehrenreich May, J. T. (2010). *Unified protocol for transdiagnostic treatment of emotional disorders: Therapist guide*. New York: Oxford University Press.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*(1), 19–21.

Baron, R. M., & Kenny, D. A. (1986). The mediator-moderator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Begg, C. B., Cho, M. K., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association*, *276*, 637–639.

Beidel, D. C., Turner, S. M., Sallee, F. R., Ammerman, R. T., Crosby, L. A., & Pathak, S. (2007). SET-C vs. Fluoxetine in the treatment of childhood social phobia. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*, 1622–1632.

Bernal, G., Bonilla, J., & Bellido, C. (1995). Ecological validity and cultural sensitivity for outcome research: Issues for the cultural adaptation and development of psychosocial treatments with Hispanics. *Journal of Abnormal Child Psychology*, *23*, 67–82.

Bernal, G., & Scharron-Del-Rio, M. R. (2001). Are empirically supported treatments valid for ethnic minorities? Toward an alternative approach for treatment research. *Cultural Diversity and Ethnic Minority Psychology*, *7*, 328–342.

Bersoff, D. M., & Bersoff, D. N. (1999). Ethical perspectives in clinical research. In P. C. Kendall,, J. Butcher, & G. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (pp. 31–55). New York, NY: Wiley.

Beutler, L. E., & Harwood, M. T. (2000). *Prescriptive therapy: A practical guide to systematic treatment selection*. New York: Oxford University Press.

Blanco, C., Olfson, M., Goodwin, R. D., Ogburn, E., Liebowitz, M. R., Nunes, E. V., & Hasin, D. S. (2008). Generalizability of clinical trial results for major depression to community samples: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Clinical Psychiatry*, *69*, 1276–1280.

Brown, R. A., Evans, M., Miller, I., Burgess, E., & Mueller, T. (1997). Cognitive-behavioral treatment for depression in alcoholism. *Journal of Consulting and Clinical Psychology*, *65*, 715–726.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.

Comer, J. S., & Kendall, P. C. (2004). A symptom-level examination of parent-child agreement in the diagnosis of anxious youths. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 878–886.

Creed, T. A., & Kendall, P. C. (2005). Therapist alliance-building behavior within a cognitive-behavioral treatment for anxiety in youth. *Journal of Consulting and Clinical Psychology*, *73*, 498–505.

Curry, J., Silva, S., Rohde, P., Ginsburg, G., Kratochvil, C., Simons, A., et al. (2011). Recovery and recurrence following treatment for adolescent major depression. *Archives of General Psychiatry*, *68*(3), 263–269.

Dawson, R., & Lavori, P. W. (2004). Placebo-free designs for evaluating new mental health treatments: The use of adaptive treatment strategies. *Statistics in Medicine*, *23*, 3249–3262.

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*, 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.

Dobson, K. S., & Hamilton, K. E. (2002). The stage model for psychotherapy manual development: A valuable tool for promoting evidence-based practice. *Clinical Psychology: Science and Practice*, *9*, 407–409.

Dobson, K. S., Hollon, S. D., Dimidjian, S., Schmaling, K. B., Kohlenberg, R. J., Gallop, R. J., et al. (2008). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. *Journal of Consulting and Clinical Psychology*, *76*, 468–477.

Dobson, K. S., & Shaw, B. (1988). The use of treatment manuals in cognitive therapy. Experience and issues. *Journal of Consulting and Clinical Psychology*, *56*, 673–682.

Flaherty, J. A., & Meaer, R. (1980). Measuring racial bias in inpatient treatment. *American Journal of Psychiatry*, *137*, 679–682.

Garfield, S. (1998). Some comments on empirically supported psychological treatments. *Journal of Consulting and Clinical Psychology*, *66*, 121–125.

Gladis, M. M., Gosch, E. A., Dishuk, N. M., & Crits-Cristoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 320–331.

Hall, G. C. N. (2001). Psychotherapy research with ethnic minorities: Empirical, ethnical, and conceptual issues. *Journal of Consulting and Clinical Psychology*, *69*, 502–510.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Heiser, N. A., Turner, S. M., Beidel, D. C., & Roberson-Nay, R. (2009). Differentiating social phobia from shyness. *Journal of Anxiety Disorders*, *23*, 469–476.

Hoagwood, K. (2002). Making the translation from research to its application: The je ne sais pas of evidence-based practices. *Clinical Psychology: Science and Practice*, *9*, 210–213.

Hollon, S. D. (1996). The efficacy and effectiveness of psychotherapy relative to medications. *American Psychologist*, *51*, 1025–1030.

Hollon, S. D., & DeRubeis, R. J. (1981). Placebo-psychotherapy combinations: Inappropriate representation of psychotherapy in drug-psychotherapy comparative trials. *Psychological Bulletin*, *90*, 467–477.

Hollon, S. D., Garber, J., & Shelton, R. C. (2005). Treatment of depression in adolescents with cognitive behavior therapy and medications: A commentary on the TADS project. *Cognitive and Behavioral Practice*, *12*, 149–155.

Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures. *Journal of Consulting and Clinical and Clinical Psychology*, *65*, 599–610.

Homma-True, R., Greene, B., Lopez, S. R., & Trimble, J. E. (1993). Ethnocultural diversity in clinical psychology. *Clinical Psychologist*, *46*, 50–63.

Hood, S., D., Potokar, J. P., Davies, S. J., Hince, D. A., Morris, K., Seddon, K. M., et al. (2010). Dopaminergic challenges in social anxiety disorder: Evidence for dopamine D3 desensitization following successful treatment with serotonergic antidepressants. *Journal of Psychopharmacology*, *24*(5), 709–716.

Huey, S. J., & Polo, A. J. (2008). Evidence-based psychosocial treatments for ethnic minority youth. *Journal of Clinical Child and Adolescent Psychology*, *37*, 262–301.

Jaccard, J., & Guilamo-Ramos, V. (2002a). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*, 130–146.

Jaccard, J., & Guilamo-Ramos, V. (2002b). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*, 278–294.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352.

Jacobson, N. S., & Hollon, S. D. (1996a). Cognitive-behavior therapy versus pharmacotherapy: Now that the jury's returned its verdict, it's time to present the rest of the evidence. *Journal of Consulting and Clinical Psychology*, *74*, 74–80.

Jacobson, N. S., & Hollon, S. D. (1996b). Prospects for future comparisons between drugs and psychotherapy: Lessons from the CBT-versus-pharmacotherapy exchange. *Journal of Consulting and Clinical Psychology*, *64*, 104–108.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects. Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307.

Jacobson, N. S., & Traux, P. (1991). Clinical significance: A statistic approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Jarrett, R. B., Vittengl, J. R., Doyle, K., & Clark, L. A. (2007). Changes in cognitive content during and following cognitive therapy for recurrent depression: Substantial and enduring, but not predictive of change in depressive symptoms. *Journal of Consulting and Clinical Psychology*, *75*, 432–446.

Jones, B., Jarvis, P., Lewis, J. A., & Ebbutt, A. F. (1996). Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal*, *313* (7048), 36–39.

Karver, M., Shirk, S., Handelsman, J. B., Fields, S., Crisp, H., Gudmundsen, G., & McMakin, D. (2008). Relationship processes in youth psychotherapy: Measuring alliance, alliance-building behaviors, and client involvement. *Journal of Emotional and Behavioral Disorders*, *16*, 15–28.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 332–339.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn and Bacon.

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.

Kendall, P. C. (1999). Introduction to the special section: Clinical Significance. *Journal of Consulting and Clinical Psychology*, *67*, 283–284.

Kendall, P. C., & Beidas, R. S. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology: Research and Practice*, *38*, 13–20.

Kendall, P. C., & Chu, B. (1999). Retrospective self-reports of therapist flexibility in a manual-based treatment for youths with anxiety disorders. *Journal of Clinical Child Psychology*, *29*, 209–220.

Kendall, P. C., & Grove, W. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment*, *10*, 147–158.

Kendall, P. C., & Hedtke, K. A. (2006). *Cognitive-behavioral therapy for anxious children* (3rd ed.). Ardmore, PA: Workbook Publishing.

Kendall, P. C., & Hollon, S. D. (1983). Calibrating therapy: Collaborative archiving of tape samples from therapy outcome trials. *Cognitive Therapy and Research*, *7*, 199–204.

Kendall, P. C., Hollon, S., Beck, A. T., Hammen, C., & Ingram, R. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy and Research*, *11*, 289–299.

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology*, *76*, 282–297.

Kendall, P. C., & Kessler, R. C. (2002). The impact of childhood psychopathology interventions on subsequent substance abuse: Policy implications, comments, and recommendations. *Journal of Consulting and Clinical Psychology*, *70*, 1303–1306.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 285–299.

Kendall, P. C., Safford, S., Flannery-Schroeder, E., & Webb, A. (2004). Child anxiety treatment: Outcomes in adolescence and impact on substance use and depression at 7.4-year follow-up. *Journal of the Consulting and Clinical Psychology*, *72*, 276–287.

Kendall, P. C., & Southam-Gerow, M. A. (1995). Issues in the transportability of treatment: The case of anxiety disorders in youth. *Journal of Consulting and Clinical Psychology*, *63*, 702–708.

Kendall, P. C., & Sugarman, A. (1997). Attrition in the treatment of childhood anxiety disorders. *Journal of Consulting and Clinical Psychology*, *65*, 883–888.

Kendall, P. C., & Suveg, C. (2008). Treatment outcome studies with children: Principles of proper practice. *Ethics and Behavior*, *18*, 215–233.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*, 990–996.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–883.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Leon, A. C., Mallinckrodt, C. H., Chuang-Stein, C., Archibald, D. G., Archer, G. E., & Chartier, K. (2006). Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology. *Biological Psychiatry*, *59*, 1001–1005.

Lin, P., Campbell, D. G., Chaney, E. F., Liu, C., Heagerty, P., Felker, B. L., et al. (2005). The influence of patient preference on depression treatment in primary care. *Annals of Behavioral Medicine*, *30*(2), 164–173.

Little, R. J. A., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lopez, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin*, *106*, 184–204.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., et al. (2002). The dodo bird verdict is alive and well—mostly. *Clinical Psychology: Science and Practice*, *9*(1), 2–12.

Marcus, S. M., Gorman, J., Shea, M. K., Lewin, D., Martinez, J., Ray, S., et al. (2007). A comparison of medication side effect reports by panic disorder patients with and without concomitant cognitive behavior therapy. *American Journal of Psychiatry*, *164*, 273–275.

Mason, M. J. (1999). A review of procedural and statistical methods for handling attrition and missing data. *Measurement and Evaluation in Counseling and Development*, *32*, 111–118.

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 the American Medical Association*, *285*, 1987–1991.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*, 445–464.

MTA Cooperative Group. (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1088–1096.

Mufson, L., Dorta, K. P., Wickramaratne, P., Nomura, Y., Olfson, M., & Weissman, M. M. (2004). A randomized effectiveness trial of interpersonal psychotherapy for depressed adolescents. *Archives of General Psychiatry*, *61*, 577–584.

Neuner, F., Onyut, P. L., Ertl, V., Odenwald, M., Schauer, E., & Elbert, T. (2008). Treatment of posttraumatic stress disorder by trained lay counselors in an African refugee settlement: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *76*, 686–694.

O'Leary, K. D., & Borkovec, T. D. (1978). Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist*, *33*, 821–830.

Olfson, M., Cherry, D., & Lewis-Fernandez, R. (2009). Racial differences in visit duration of outpatient psychiatric visits. *Archives of General Psychiatry*, *66*, 214–221.

Pediatric OCD Treatment Study (POTS) Team. (2004). Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive-compulsive disorder: The Pediatric OCD Treatment Study (POTS) randomized controlled trial. *Journal of the American Medical Association*, *292*, 1969–1976.

Pelham, W. E., Jr., Gnagy, E. M., Greiner, A. R., Hoza, B., Hinshaw, S. P., Swanson, J. M., et al. (2000). Behavioral versus behavioral and psychopharmacological treatment in ADHD children attending a summer treatment program. *Journal of Abnormal Child Psychology*, *28*, 507–525.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Reis, B. F., & Brown, L. G. (2006). Preventing therapy dropout in the real world: The clinical utility of videotape preparation and client estimate of treatment duration. *Professional Psychology: Research and Practice*, *37*, 311–316.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., et al. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials*, *25*(1), 119–142.

Sachs, G. S., Thase, M. E., Otto, M. W., Bauer, M., Miklowitz, D., Wisniewski, S. R., et al. (2003). Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biological Psychiatry*, *53*(11), 1028–1042.

Shirk, S. R., Gudmundsen, G., Kaplinski, H., & McMakin, D. L. (2008). Alliance and outcome in cognitive-behavioral therapy for adolescent depression. *Journal of Clinical Child and Adolescent Psychology*, *37*, 631–639.

Silverman, W. K., Kurtines, W. M., & Hoagwood, K. (2004). Research progress on effectiveness, transportability, and dissemination of empirically supported treatments: Integrating theory and research. *Clinical Psychology: Science and Practice*, *11*, 295–299.

Snowden, L. R. (2003). Bias in mental health assessment and intervention: Theory and evidence. *American Journal of Public Health*, *93*, 239–243.

Southam-Gerow, M. A., Ringeisen, H. L., & Sherrill, J. T. (2006). Integrating interventions and services research: Progress and prospects. *Clinical Psychology: Science and Practice*, *13*, 1–8.

Sue, S. (1998). In search of cultural competence in psychotherapy and counseling. *American Psychologist*, *53*, 440–448.

Suveg, C., Comer, J. S., Furr, J. M., & Kendall, P. C. (2006). Adapting manualized CBT for a cognitively delayed child with multiple anxiety disorders. *Clinical Case Studies*, *5*, 488–510.

Sweeney, M., Robins, M., Ruberu, M., & Jones, J. (2005). African-American and Latino families in TADS: Recruitment and treatment considerations. *Cognitive and Behavioral Practice*, *12*, 221–229.

Treadwell, K., Flannery-Schroeder, E. C., & Kendall, P. C. (1994). Ethnicity and gender in a sample of clinic-referred anxious children: Adaptive functioning, diagnostic status, and treatment outcome. *Journal of Anxiety Disorders*, 9, 373–384.

Treatment for Adolescents with Depression Study (TADS) Team. (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents with Depression Study (TADS) randomized controlled trial. *Journal of the American Medical Association*, *292*, 807–820.

Vanable, P. A., Carey, M. P., Carey, K. B., & Maisto, S. A. (2002). Predictors of participation and attrition in a health promotion study involving psychiatric outpatients. *Journal of Consulting and Clinical Psychology*, *70*, 362–368.

Walders, N., & Drotar, D. (2000). Understanding cultural and ethnic influences in research with child clinical and pediatric psychology populations. In D. Drotar (Ed.), *Handbook of research in pediatric and clinical child psychology* (pp. 165–188). New York: Springer.

Walkup, J. T., Albano, A. M., Piacentini, J., Birmaher, B., Compton, S. N., et al. (2008) Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine*, *359*, 1–14.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61*, 620–630.

Weersing, R. V., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology*, *70*(2), 299–310.

Weisz, J., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*, 688–701.

Weisz, J. R., Thurber, C. A., Sweeney, L., Proffitt, V. D., & LeGagnoux, G. L. (1997). Brief treatment of mild-to-moderate child depression using primary and secondary control enhancement training. *Journal of Consulting and Clinical Psychology*, *65*(4), 703–707.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist*, *47*, 1578–1585.

Westbrook, D., & Kirk, J. (2007). The clinical effectiveness of cognitive behaviour therapy: Outcome for a large sample of adults treated in routine practice. *Behaviour Research and Therapy*, *43*, 1243–1261.

Westen, D., Novotny, C., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, *130*, 631–663.

Wilson, G. T. (1995). Empirically validated treatments as a basis for clinical practice: Problems and prospects. In S. C. Hayes, V. M. Follette, R. D. Dawes & K. Grady (Eds.), *Scientific standards of psychological practice: Issues and recommendations* (pp. 163–196). Reno, NV: Context Press.

Yeh, M., McCabe, K., Hough, R. L., Dupuis, D., & Hazen, A. (2003). Racial and ethnic differences in parental endorsement of barriers to mental health services in youth. *Mental Health Services Research*, *5*, 65–77.

CHAPTER

5

Dissemination and Implementation Science: Research Models and Methods

Rinad S. Beidas, Tara Mehta, Marc Atkins, Bonnie Solomon, *and* Jenna Merz

Abstract

Dissemination and implementation (DI) science has grown exponentially in the past decade. This chapter reviews and discusses the research methodology pertinent to empirical DI inquiry within mental health services research. This chapter (a) reviews models of DI science, (b) presents and discusses design, variables, and measures relevant to DI processes, and (c) offers recommendations for future research.

Key Words: Dissemination, implementation, research methods, mental health services research

Introduction

Using the specific criteria for "empirically supported treatments" (Chambless & Hollon, 1998), efficacious psychosocial treatments have been identified for mental health and substance abuse, and national accrediting bodies (e.g., American Psychological Association [APA]) have recommended the use of such treatments, a practice that is often referred to as *evidence-based practice* (EBP; APA, 2005). However, uptake of EBP is a slow process, with some suggesting that the translation of new research findings into clinical practice can take over a decade and a half (Green, Ottoson, Garcia, & Hiatt, 2009). Given the emphasis on dissemination and implementation of research innovation, a number of recent efforts have endeavored to ensure that EBP is disseminated to and implemented within the community (McHugh & Barlow, 2010). For example, the United States Veterans Administration Quality Enhancement Research Initiative and the United Kingdom's Improving Access to Psychological Therapies are examples of international efforts to enact large-scale systemic change in the provision of EBP.

Part of the impetus for the EBP movement in mental health services in the United States was a 1995 task force report initiated by the Division of Clinical Psychology (Division 12) of the APA (Chambless & Hollon, 1998). The initial report identified empirically supported psychosocial treatments for adults and also highlighted the lack of empirical support for many interventions. Since the initial report, debate with regard to the provision of EBP in clinical practice has ensued, but the movement has gained solid footing. Efforts to expand the use of EBP have encouraged the rethinking of community mental health practice, akin to the movement within evidence-based medicine (Torrey, Finnerty, Evans, & Wyzik, 2003).

Given the different terms used within the area of dissemination and implementation (DI) research (Beidas & Kendall, 2010; Special Issue of the *Journal of Consulting and Clinical Psychology*, Kendall & Chambless, 1998; Rakovshik & Mcmanus, 2010), operational definitions are provided. EBP refers here to the provision of psychosocial treatments supported by the best scientific evidence while also taking into account clinical experience and client preference (APA, 2005).

Empirically supported treatments refer here to specific psychological interventions that have been evaluated scientifically (e.g., a randomized controlled trial [RCT]) and independently replicated with a delineated population (Chambless & Hollon, 1998). DI science includes the purposeful distribution of relevant information and materials to therapists (i.e., dissemination) and the adoption and integration of EBP into practice (i.e., implementation; Lomas, 1993). Dissemination and implementation are best initiated together in that both need to occur in order to influence systemic change (Proctor et al., 2009).

This relatively nascent field of study has yet to develop a "gold-standard" set of research methods specific to DI processes. Nevertheless, this chapter reviews relevant research methodology pertinent to research questions within this area. The chapter (a) reviews models of DI science, (b) presents and discusses relevant research methods (i.e., design, variables, and measures), and (c) offers recommendations for future research.

Research Methods

Models

A number of models[1] exist that are specific to DI science (e.g., Consolidated Framework for Implementation Research; Damschroder et al., 2009) or have been applied from other areas (e.g., Diffusion of Innovation; Rogers, 1995) that are salient. When considering models, it is important to consider model typology and the need for multiple models to explain DI processes (Damschroder, 2011). Impact models are explanatory in that they describe DI hypotheses and assumptions, including causes, effects, and factors (i.e., the "what"), whereas process models emphasize the actual implementation process (i.e., the "how to"; Grol, Bosch, Hulscher, Eccles, & Wensing, 2007). Below, relevant models are described. First, we present heuristic models that can guide study conceptualization, and then we present models that are more specific to various DI questions, including models that emphasize individual practitioners and social and organizational processes. See Table 5.1 for a comparison of DI models and features.

COMPREHENSIVE MODELS

Models included within this section are comprehensive and ecological in nature in that they include individual, organizational, and systemic processes. These models function largely as guiding heuristics when designing DI studies and include Promoting Action on Research Implementation in Health Services (PARiHS; Kitson, Harvey, & McCormack, 1998); Reach, Efficacy, Adoption, Implementation, and Maintenance (RE-AIM; Glasgow, Vogt, & Boles, 1999); Stages of Implementation and Core Implementation Components (Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005); the Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009); the Practical, Robust Implementation, and Sustainability Model (PRISM) (Feldstein & Glasgow, 2008); and a Conceptual Model of Implementation Research (Proctor et al., 2009).

PARiHS

The PARiHS framework has been put forth as a practical heuristic to understand the process of implementation (Kitson et al., 2008). The use of the PARiHS model is twofold, "as a diagnostic and evaluative tool to successfully implement evidence into practice, and by practitioners and researchers to evaluate such activity" (Kitson et al., 2008).

The framework posits three interactive components: evidence (E), context (C), and facilitation (F). E refers to knowledge, C refers to the system within which implementation occurs, and F refers to support of the implementation process. Successful implementation depends on the interrelationship between E, C, and F (Kitson et al., 1998). The PARiHS model emphasizes that (a) evidence is composed of "codified and non-codified source of knowledge," which includes research, clinical experience, patient preferences, and local information, (b) implementing evidence in practice is a team effort that must balance a dialectic between new and old, (c) certain settings are more conducive to implementation of new evidence than others, such as those that have evaluation and feedback in place, and (d) facilitation is necessary for implementation success (Kitson et al., 2008). Initial support exists around the model (e.g., facilitation; Kauth et al., 2010), although there is the need for prospective study (Helfrich et al., 2010).

RE-AIM

The RE-AIM framework is another model that can aid in the planning and conducting of DI studies. RE-AIM evaluates the public health impact of an intervention as a function of the following five factors: reach, efficacy, adoption, implementation, and maintenance. This model is consistent with a systems-based social ecological framework (Glasgow et al., 1999).

Table 5.1 A Comparison of Key Features Across Key Dissemination and Implementation (DI) Models

Model	Comprehensive/ Ecological	Emphasizes Individual Practitioner	Emphasizes Social and Organizational Processes	Emphasizes Feedback Loops	Emphasizes Process of Implementation	Emphasizes Sustainability	Considers Role of Research–Community Partnerships	Emphasizes Key Opinion Leaders	Empirically Evaluated
	Model Features								
PARiHS	■	■	■		■				■
RE-AIM	■	■	■		■	■			■
SI/CIC	■	■	■		■	■			■
CFIR	■	■	■		■				
PRISM	■	■	■		■	■	■		
CMIR	■	■	■		■				
Stetler	■	■	■						
TPB		■							■
DOI		■	■		■			■	■
ARC	■	■	■		■		■	■	■
CID			■		■	■			

Note: PARiHS = Promoting Action on Research Implementation in Health Services (Kitson et al., 1998); RE-AIM = Reach, Efficacy, Adoption, Implementation, and Maintenance (Glasgow et al., 1999); SI/CIC = Stages of Implementation and Core Implementation Components (Fixsen et al., 2005); CFIR = Consolidated Framework for Implementation Research (Damschroder et al., 2009); PRISM = Practical, Robust Implementation, and Sustainability Model (Feldstein & Glasgow, 2008); CMIR = Conceptual Model of Implementation Research (Proctor et al., 2009); Stetler model (Stetler, 2001); TPB = Theory of Planned Behavior (Ajzen, 1988; 1991); DOI = Diffusion of Innovation (Rogers, 1995); ARC = Availability, Responsiveness, and Continuity model (Glisson & Schoenwald, 2005); CID = Clinic/Community Intervention Development Model (Hoagwood et al. 2002).

■ = Feature characterizes model

Reach refers to "the percentage and risk characteristics of persons who receive or are affected by a policy or program," whereas efficacy refers to positive and negative health outcomes (i.e., biological, behavioral, and patient-centered) following implementation of an intervention (Glasgow et al., 1999, p. 1323). Both reach and efficacy address individual-level variables. Adoption refers to the number of settings that choose to implement a particular intervention, whereas implementation refers to "the extent to which a program is delivered as intended" (Glasgow et al., 1999, p. 1323). Both adoption and implementation are organizational-level variables. Maintenance refers to the extent to which an intervention becomes a routine part of the culture of a context (i.e., sustainability). Maintenance is both an individual- and organizational-level variable. Each of the five factors can be scored from 0 to 100, with the total score representing the public health impact of a particular intervention. Interventions, such as various EBPs, can be scored on each dimension and plotted and compared to one another.

Over 100 studies have been completed using RE-AIM as an organizing heuristic since it was published in 1999, but the authors state that it has not been validated because it is a guiding framework rather than model or theory (http://www.re-aim.org/about_re-aim/FAQ/index.html). No literature reviews of RE-AIM–guided studies exist to our knowledge.

Stages of Implementation and Core Implementation Components

Fixsen and colleagues have provided two key conceptual models for understanding implementation processes (Fixsen, Blase, Naoom, & Wallace, 2009; Fixsen et al., 2005). The recursive and nonlinear stages of implementation include exploration, installation, initial implementation, full implementation, innovation, and sustainability (Fixsen

et al., 2005). Fixsen and colleagues (2009) suggest that "the stages of implementation can be thought of as components of a tight circle with two-headed arrows from each to every other component" (Fixsen et al., 2009, p. 534).

Based upon a review of successful programs, a number of core components were proposed within the stages of implementation. These core implementation components include: staff selection, preservice and in-service training, ongoing coaching and consultation, staff evaluation, decision support data systems, facilitative administrative support, and systems interventions (Fixsen et al., 2005). These components are both integrated and compensatory in that they work together and compensate for strengths and weaknesses to result in optimal outcomes. Core implementation components work in tandem with effective programs (Fixsen et al., 2005).

Given the integrated and compensatory nature of the core implementation components, an adjustment of one necessarily influences the others. Importantly, feedback loops must be built into implementation programs that allow for such natural corrections. These core implementation components provide a blueprint for implementation research design (Fixsen et al., 2009). Although this model focuses on clinician behavior as the emphasized outcome variable, systemic variables and patient outcomes are also included, making it a comprehensive model of implementation processes.

CFIR

The CFIR is a metatheoretical synthesis of the major models emerging from implementation science (Damschroder et al., 2009). CFIR does not specify hypotheses, relationships, or levels but rather distills models and theories into core components, creating an overarching ecological framework that can be applied to various DI research studies. CFIR has five major domains that reflect the structure of other widely cited implementation theories (e.g., Fixsen et al., 2005; Kitson et al., 1998): intervention characteristics, outer setting, inner setting, individual characteristics, and the implementation process.

Intervention characteristics are important in DI, particularly the core (i.e., essential elements) and peripheral (i.e., adaptable elements) components. Other important intervention characteristics include intervention source, stakeholder perception of the evidence for the intervention, stakeholder perception of the advantage of implementing the intervention, adaptability of the intervention for a particular setting, feasibility of implementing a pilot, complexity of the intervention, design quality and packaging, and cost.

The outer setting refers to the "economic, political, and social context within which an organization resides" (Damschroder et al., 2009, p. 57). Specifically, the outer setting concerns patient needs for the intervention, cosmopolitanism (i.e., the social network of the organization), peer pressure to implement the intervention, and external incentives to implement. The inner setting refers to the "structural, political, and cultural contexts through which the implementation process will proceed" (Damschroder et al., 2009, p. 57). These include structural characteristics (e.g., social organization of the agency, age, maturity, size), social networks and communication, culture, and implementation climate. The outer setting can influence implementation and may be mediated through modifications of the inner setting, and the two areas can be overlapping and dynamic (Damschroder et al., 2009).

Individual characteristics refer to stakeholders involved with the process of implementation. This framework views stakeholders as active seekers of innovation rather than passive vessels of information (Greenhalgh, Robert, Macfarlane, Bate, & Kyriakidou, 2004). Constructs within this domain include knowledge and beliefs about the intervention, self-efficacy with regard to use of the intervention, individual stage of change, individual identification with the organization, and other individual attributes. Finally, the implementation process here refers to four activities: planning, engaging, executing, and reflecting and evaluating. Empirical validation for the CFIR model is currently ongoing.

PRISM

The PRISM model (Feldstein & Glasgow, 2008) represents another comprehensive ecological model that integrates across existing DI frameworks (e.g., PARiHS, RE-AIM) to offer a guiding heuristic in DI study design. PRISM is comprehensive in that it "considers how the program or intervention design, the external environment, the implementation and sustainability infrastructure, and the recipients influence program adoption, implementation, and maintenance" (Feldstein & Glasgow, 2008, p. 230).

The first element of the PRISM model considers the perspectives of the organization and consumers with regard to the intervention. Organizational characteristics are investigated at three levels (leadership, management, and front-line staff); the authors

recommend considering how the intervention will be perceived by the organization and staff members. For example, readiness for change, program usability, and alignment with organizational mission are a few issues to address. With regard to taking the consumer perspective, PRISM recommends considering how an intervention will be received by consumers, such as burden associated with the intervention and the provision of consumer feedback.

The second element of PRISM focuses on organizational and consumer characteristics. Important organizational characteristics include the financial and structural history of an organization as well as management support. Consumer characteristics to consider include demographics, disease burden, and knowledge and beliefs. Relatedly, the third element considers characteristics of the external environment relevant to DI efforts, which "may be some of the most powerful predictors of success" (Feldstein & Glasgow, 2008, p. 237). The external environment refers to motivating variables such as payer satisfaction, competition, regulatory environment, payment, and community resources.

The fourth element of the PRISM model refers to the infrastructure present to support implementation and sustainability. The authors recommend that for implementation to be successful, plans for sustainability must be integrated into DI efforts from the very beginning. Specific variables to consider within this element include adopter training and support, adaptable protocols and procedures, and facilitation of sharing best practices.

The unique contributions of the PRISM model lie in the integration of various DI models and focus on integrating concepts not included in previous models: (a) perspectives and characteristics of organizational workers at three levels (leadership, management, and staff), (b) partnerships between researchers and those doing the implementation, and (c) planning for sustainability from the beginning. Additionally, the authors provide a useful set of questions to ask at each level of the PRISM model when designing a research project (see Feldstein & Glasgow, 2008).

Conceptual Model of Implementation Research

Proctor and colleagues (2009) proposed a conceptual model of implementation research that integrates across relevant theories and underscores the types of outcomes to consider in DI research. Their model assumes nested levels (policy, organization, group, individual) that integrate quality improvement, implementation processes, and outcomes. The model posits two required components: evidence-based intervention strategies (i.e., EBP) and evidence-based implementation strategies (i.e., systems environment, organizational, group/learning, supervision, individual providers/consumers). Unique to this model, three interrelated outcomes are specified: implementation (e.g., feasibility, fidelity), service (e.g., effectiveness, safety), and client (e.g., symptoms) outcomes.

MODELS THAT EMPHASIZE INDIVIDUAL PRACTITIONERS

Moving beyond heuristic models, we describe models that specify various components of DI processes. Models included within this section emphasize individual practitioners and include the Stetler model (Stetler, 2001) and Theory of Planned Behavior (Ajzen, 1988, 1991).

Stetler Model

The Stetler model (Stetler, 2001) emerges from the nursing literature and focuses on how the individual practitioner can use research information in the provision of EBP. The linear model is "a series of critical-thinking steps designed to buffer the potential barriers to objective, appropriate, and effective utilization of research findings" (Stetler, 2001). The unit of emphasis is the individual's appropriate use of research findings.

The Stetler model has been updated and refined a number of times (Stetler, 2001) and comprises five main stages: (a) preparation, (b) validation, (c) comparative evaluation/decision making, (d) translation/application, and (e) evaluation. During preparation, the practitioner identifies a potential high-priority problem, considers the need to form a team or other internal and/or external factors, and seeks systematic reviews and empirical evidence relevant to the problem. During validation, the practitioner rates the quality of evidence and rejects noncredible sources. During comparative evaluation/decision making, the practitioner synthesizes findings across empirical sources, evaluates the feasibility and fit of current practices, and makes a decision about the use of evidence in the problem identified. During translation/application, the evidence is used with care to ensure that application does not go beyond the evidence. Additionally during this stage, a concerted effort to include dissemination and change strategies is necessary. During evaluation, outcomes from the implementation of the evidence are assessed, including both formal and informal evaluation and cost/benefit analyses. Both formative and

summative evaluations are to be included (Stetler, 2001).

Theory of Planned Behavior

Theory of Planned Behavior (TPB; Ajzen, 1988; 1991) can be used to understand the behavior of the individual practitioner within DI efforts. From the perspective of TPB, behavior is determined by an individual's intention to perform a given behavior. Intentions are a function of attitudes toward the behavior, subjective norms, and perceived control. This theory has received great attention in other areas of psychology and is empirically supported (Armitage & Conner, 2001) but has only recently been applied to DI processes.

In one recent study, clinicians were randomly assigned to one of two continuing education workshops: a TPB-informed workshop and a standard continuing-education workshop. Outcomes included clinician intentions and behavior in the usage of an assessment tool. The key manipulation in the TPB-informed workshop was an elicitation exercise to gather participant attitudes, social norms, and perceived control. Findings were supportive in that participants demonstrated both higher intentions and higher implementation rates in the use of the assessment tool (Casper, 2007). This model can be used to guide the design of studies hoping to influence behavior change at the individual practitioner level.

MODELS THAT EMPHASIZE SOCIAL AND ORGANIZATIONAL PROCESSES

Models within this section emphasize the social nature of DI and the importance of organizational context and include Diffusion of Innovation (Rogers, 1995), the Availability, Responsiveness, and Continuity model (Glisson & Schoenwald, 2005), and the Clinic/Community Intervention Development Model (Hoagwood, Burns, & Weisz, 2002).

Diffusion of Innovation

The Diffusion of Innovation (DOI) framework (Rogers, 1995) has been widely used and cited within the field of DI science as an integral framework. DOI has been empirically applied across a number of fields, such as agriculture and health sciences (Green et al., 2009). The tenets of DOI are outlined in Rogers' book, *Diffusion of Innovations*, which was revised to its fifth edition before Rogers' death in 2004. Over 5,000 studies have been conducted on DOI, and a new one is published approximately daily (Rogers, 2004).

Rogers defined diffusion as "the process through which an innovation, defined as an idea perceived as new, spreads via certain communication channels over time among the members of a social system" (Rogers, 2004, p. 13). Diffusion can be conceptualized as both a type of communication and of social change that occurs over time (Haider & Kreps, 2004). Adoption of innovation is contingent upon five characteristics: relative advantage, compatibility, complexity, trialability, and observability (Rogers, 1995). Relative advantage refers to whether or not use of an innovation will confer advantage to the individual (e.g., improve job performance, increase compensation). Compatibility is the extent to which an innovation is consistent with the individual's set of values and needs. Complexity refers to how easily an innovation can be learned and used. Trialability is the extent to which an innovation can be tested on a small scale to evaluate efficacy. Observability describes the positive outcomes that are engendered by implementation of an innovation.

Irrespective of innovation characteristics, DOI theory suggests that innovations are adopted according to a five-step temporal process of Innovation-Decision: knowledge, persuasion, decision, implementation, and confirmation. Knowledge refers to an individual learning of an innovation, whereas persuasion refers to attitude formation about an innovation. Decision occurs when a person decides to adopt or reject an innovation. Implementation refers here to when an individual uses an innovation, whereas confirmation refers to an individual seeking reinforcement about the decision to implement an innovation. Decisions to adopt an innovation are recursive, meaning that an individual can reject an innovation at first while adopting it later (Lovejoy, Demireva, Grayson, & McNamara, 2009). Rogers (2004) describes the diffusion of innovation as following an S-shaped curve where innovation adoption begins at a slow rate (i.e., early adopters; first 16%) but reaches a tipping point when adoption accelerates rapidly (i.e., early and late majority; 68%) and then decreases again (i.e., laggards; last 16%). The tipping point, or threshold of program utilizers, occurs when approximately 25% of the social network become utilizers (Valente & Davis, 1999). A well-known and practical application of DOI includes key opinion leaders, a small group of influential early adopters who make it more likely that innovation will spread within a social network (Valente & Davis, 1999); this theory has been supported in mental health services research (Atkins et al., 2008).

DOI has been influential in DI science. The field has taken into account characteristics of innovations and the innovation-decision process within a social context when designing DI research. DOI has been applied to understanding how to bridge the gap between research and clinical practice within various psychosocial interventions and treatment populations (e.g., autism; Dingfelder & Mandell, 2010).

Availability, Responsiveness, and Continuity Model

The Availability, Responsiveness, and Continuity (ARC) organizational and community model is specific to mental health services research and is based upon three key assumptions: (a) the implementation of EBP is both a social and technical process, (b) mental health services are embedded in layers of context, including practitioner, organization, and community, and (c) effectiveness is related to how well the social context can support the objectives of the EBP (Glisson & Schoenwald, 2005). ARC aims to improve the fit between the social context and EBP through intervening at the organizational and interorganizational domain levels. The organizational level refers to the needs of mental health practitioners, and ARC involves such providers in organizational processes and policies. The emphasis on interorganizational domain level within ARC allows for the formation of partnerships among practitioners, organizational opinion leaders, and community stakeholders with the shared goal of ameliorating identified problems in a community through a particular EBP (Glisson & Schoenwald, 2005).

Within the ARC model, a key component includes an ARC change agent who "works with an interorganizational domain (e.g., juvenile court, school system, law enforcement, business group, churches) at several levels (e.g., community, organization, individual) around a shared concern (e.g., reducing adolescent delinquent behavior)" (Glisson & Schoenwald, 2005, p. 248). This individual works at the community level by helping form a group to support an EBP for a selected population, at the organizational level by providing support in the delivery of EBP, and at the individual level to develop individual partnerships with key opinion leaders. Change agents provide technical information, empirical evidence, evaluation of outcomes, and support during times of conflict. In other words, the role of the change agent is to serve as a bridge between those disseminating and those implementing the EBP (Glisson & Schoenwald, 2005). An especially clear and relevant application of ARC is described in a recent study that improved DI efforts of multisystemic therapy into poor rural communities (Glisson et al., 2010).

Clinic/Community Intervention Development Model

Hoagwood, Burns, and Weisz (2002) proposed the Clinic/Community Intervention Development (CID) model for community deployment efforts of EBP for youth mental health. The CID model allows DI researchers to understand factors associated with sustainable services, including why and how services work in practice settings. The CID model comprises eight steps. Steps 1 through 6 involve efficacy to effectiveness with emphasis on single case applications in practice settings, a limited effectiveness study to pilot the intervention in real-world practice settings, followed by a full effectiveness study. Steps 7 and 8 are specific to DI processes. Step 7 calls for a series of studies to assess goodness of fit with practice settings, whereas Step 8 focuses on going to scale by engaging in dissemination research in multiple organizational settings.

CID is put forth as a model "for speeding up the process of developing scientifically valid and effective services within the crucible of practice settings" (Hoagwood et al., 2002, p. 337). A strength of the model is that it is externally valid given its emphasis on components, adaptations, and moderators and mediators. Additionally, the model calls for innovative thinking as well as new research models to assess goodness of fit and criteria to determine when a program is ready to go to scale.

SUMMARY

As evident from this review, the sheer number of possible DI models to consider when designing a research question can be quite daunting. Each model presented has strengths and limitations, and none of the models offered covers all of the content areas relevant to DI science (see Table 5.1). One clear limitation is that many of these newly derived theoretical models have not yet been subjected to rigorous scientific evaluation. Despite these limitations, we recommend that all DI-related questions be theoretically driven. When designing a research question, first identify a relevant model that can guide the construction of research design in order to provide meaningful contributions to the field. Our bias and recommendation is toward comprehensive ecological models that take into account the contextual aspects of DI processes as the underlying framework. However, when examining certain processes (e.g., attitudes), it can be helpful to select specific models that can lead to testable hypotheses.

For example, one might select a heuristic model such as the CFIR (Damschroder et al., 2009) when considering which constructs to focus on in a DI study and then select a more specific model based on the study question (e.g., training and attitudes; TPB, Ajzen, 1988, 1991).

We concur with Damschroder's (2011) suggestions of the following steps when selecting models: consider (a) the nature of the model (i.e., process vs. impact, context, discipline), (b) level of application (e.g., individual, organization), (c) available evidence, and (d) which model has the greatest potential for adding to the literature. Importantly, it is likely that more than one model will be needed when designing complex DI studies. Furthermore, after aggregating results, it is important to consider how the results fit back in with the original model(s) selected with regard to validation of the mode and necessary refinements (Damschroder, 2011).

Research Design

The most relevant research designs for DI studies are provided and discussed. Although all of the research methods addressed within this book may be appropriate in the design of DI studies, given the size and complexity of such studies, we focus on designs that are particularly salient to DI: experimental designs, quasi-experimental designs, and qualitative methodology.

EXPERIMENTAL DESIGNS

Randomized Controlled Trials

A full discussion of randomized controlled trials (RCTs) is beyond the scope of this chapter (see Kendall & Comer, 2011); however, RCT designs are often used in DI studies and merit mention (e.g., Miller, Yahne, Moyers, Martinez, & Pirritano, 2004; Sholomskas et al., 2005). The main strength of RCTs involves the use of random assignment to rule out selection bias, which allows for differences in outcomes between conditions to be explained by the experimental manipulation rather than group differences (Song & Herman, 2010). RCTs are often considered the gold-standard research design.

Much has been written about the use of RCTs in DI research. Some researchers have suggested that limitations exist to RCTs in their application to DI studies (e.g., Atkins, Frazier, & Cappella, 2006; Carroll & Rounsaville, 2003). Such limitations include tightly controlled settings, homogenous participants (although some research suggests this is overstated; see Stirman, DeRubeis, Crits-Cristoph, & Brody, 2003), resource-intensiveness, and delay in application of findings to practice (Atkins et al., 2006; Carroll & Rounsaville, 2003). In addition, DI trials often operate at a larger system level, requiring that the unit of randomization be at the system level (e.g., agencies, schools, classrooms, work settings). Thus, the sample needed to have adequate power to detect differences beyond chance may be beyond the capacity of many DI trials.

Clinical Equipoise

One option for augmenting traditional RCT designs for DI research in a flexible manner comes from clinical equipoise. Freedman (1987) suggested the use of clinical equipoise in RCTs. The criterion for clinical equipoise is met if there is genuine uncertainty within the practice community about a particular intervention. Statistical procedures have been developed that allow for balancing the principle of clinical equipoise with randomization (i.e., equipoise-stratified randomized design; Lavori et al., 2001).

For example, in the case of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) research trial (Rush, 2001), a patient and a service provider might agree that all treatments possible after a failed trial of citalopram are roughly equivalent (clinical equipoise). Using an equipoise-stratified randomized design allows the clinician and patient to judge what the best treatment option might be based on patient preferences, which can then be statistically controlled by using chosen treatment option as a prerandomization factor (see Lavori et al., 2001, for a detailed description). From a DI perspective, equipoise offers an advance over the constraints typically imposed on participants and settings in RCTs (e.g., West et al., 2008). The concept of equipoise has been integrated into Sequential Multiple Assignment Randomized Trial (SMART) designs, which allow for patient and provider preference while maintaining the use of randomization and the rigor of RCTs (Landsverk, Brown, Rolls Reutz, Palinkas, & Horwitz, 2010) (see also Chapter 4 in this volume). SMART designs make possible experimental investigation of the treatment choices made by patients and providers by using randomization strategies that account for choice.

Standardization

Another way to consider how to augment RCTs for DI research is to determine which components of the intervention require standardization (Hawe,

Shiell, & Riley, 2004). A complex intervention refers to an intervention that cannot be simply reduced into component parts to understand the whole (i.e., component analysis; Hawe et al., 2004). However, because an intervention is complex does not mean that an RCT is not appropriate—the question lies in what part of the intervention is standardized. Standardization as it is conceptualized within a traditional RCT suggests that components of the intervention are the same across different sites. Hawe and colleagues (2004) suggest an alternative perspective to standardization: "rather than defining the components of the intervention as standard—for example, the information kit, the counseling intervention, the workshops—what should be defined as standard are the steps in the change process that the elements are purporting to facilitate or the key functions that they are meant to have" (Hawe et al., 2004, p. 1562).

Pragmatically, this means that the form can be adapted while process and function remain standardized. What is varied becomes the form of the intervention in different contexts. For example, to train providers about treatment of anxiety, the traditional way to conduct an RCT would be to standardize training methods across sites. Live group instruction might be compared to computer-guided instruction. In each case, the information provided would be the same and therefore the results would relate to which type of training was superior for the majority of participants. Alternatively, one could standardize the *function* by providing supervisors in an organization with the materials necessary to create training programs that are tailored to the specific setting. In this case, intervention integrity would not relate to typical quality assurance efforts (i.e., did trainer follow specific protocol); rather, it would be related to whether the training developed within each context provided information consistent with the theory or principles underlying the change process. This effort could result in improved effectiveness of DI efforts (Hawe et al., 2004).

Practical Clinical Trials

Practical clinical trials (PCTs; also known as pragmatic clinical trials) have been recommended as an alternative to traditional RCTs (Tunis, Stryer, & Clancy, 2003) and are specifically relevant to effectiveness studies, the mainstay of DI research. PCTs are designed to provide the information necessary to make decisions about best-care practices in routine clinical settings. Tunis and colleagues (2003) describe the distinctive features of PCTs in that "they select clinically relevant interventions to compare, include a diverse population of study participants, recruit participants from a variety of practice settings, and collect data on a broad range of health outcomes" (p. 1626).

March and colleagues (2005) suggested that there are eight defining principles of PCTs: (a) research questions that are of public health interest and clinically relevant, (b) they are performed in usual care settings, (c) power is sufficient to identify small to medium effects, (d) randomization is included, (e) randomization depends on the principle of uncertainty/clinical equipoise, (f) outcomes are simple and clinically relevant, (g) interventions map onto best clinical practice, and (h) research burden is minimized. PCTs are well suited to answer questions related to intervention effectiveness as well as which treatment works best for which patients depending on their characteristics (March et al., 2005).

PCTs are similar to effectiveness designs in that they aim to provide information to decision makers about whether or not interventions work in routine clinical care settings. Questions best answered by this design include the overall effectiveness of a particular intervention in routine settings and include heterogeneous patient populations necessitating larger sample sizes. Outcomes must include evidence that is relevant to everyday policymakers, such as quality of life and the cost-effectiveness of interventions (Macpherson, 2004).

The main strength of PCTs and their relevance to DI research lies in their emphasis on understanding whether or not interventions can be effective in real-world settings. In other words, these designs are heavy on external validity and ecological evidence and provide evidence for decision makers regarding which interventions to recommend. Such trials have been used effectively in medicine and psychiatry (March et al., 2005). Limitations to PCTs include that they are very costly, need a resource-intensive infrastructure to succeed (March et al., 2005; Tunis et al., 2003), require close collaborations between the research team and practice sites, and may be more reflective of agency priorities than researcher priorities (e.g., symptom reduction may not be the primary outcome measure but rather improved functioning in setting). However, recent advances in electronic health records make it more feasible to realize the potential of such designs in the future (March, 2011).

Adaptive Clinical Trials

Adaptive clinical trials are another alternative to RCTs and are flexible in that they plan for the

possibility of reactive changes to study design and/or statistical procedures as the study progresses based upon review of interim data (Chow & Chang, 2008). That is, an adaptive design can be defined as "a design that allows adaptations to trial and/or statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial" (Chow & Chang, 2008). A number of adaptive design strategies exist (see Chow & Chang, 2008, for review). One that stands out as being particularly salient to DI processes includes adaptive treatment switching. This design allows researchers to switch a participant from one group to another based on lack of efficacy. For example, a patient assigned to usual care could be switched to an EBP if usual care is not effective. Bayesian analytic approaches that rely on probability theory are especially appropriate statistical analyses for these designs (Luce et al., 2009).

Although these designs have the advantage of allowing for flexibility to accommodate policy-related questions, they provide challenges to fidelity assessment and there are as yet no clear guidelines for the appropriate use of adaptive clinical trial designs (Chow & Chang, 2008).

Hybrid Models

Hybrid models have been recommended to capitalize on the best of efficacy and effectiveness methodologies (Atkins et al., 2006; Carroll & Rounsaville, 2003). Carroll and Rounsaville (2003) proposed a hybrid model that retains the methodological rigor of RCTs but adds additional components of traditional effectiveness research. In addition to the typical features of an RCT meant to protect internal validity (e.g., random assignment, blind assessment of outcomes, fidelity monitoring), the authors suggest that the following components be integrated into the design to balance external validity and make RCTs more appropriate for DI research: enhanced diversity in patients and settings, attention to training issues, evaluation of cost effectiveness, and assessment of patient and provider satisfaction. These recommendations have been feasibly integrated into DI RCTs. For example, one study feasibly balanced features of efficacy research (e.g., randomization, rigorous assessment) and effectiveness research (e.g., few exclusion criteria, completed in naturalistic setting; Dimeff et al., 2009). Other important recommendations when considering how to adapt RCT methodology for DI research include understanding organizational context and including a "systematic and iterative approach to study development" (Atkins et al., 2006, p. 107). This allows for flexible research design and "ongoing interaction between researcher- and context-driven information at various information points in a project" (Atkins et al., 2006, p. 107).

QUASI-EXPERIMENTAL DESIGNS

Single-Case Time-Series Intervention

Given the emphasis within the psychological literature on RCTs, single-case time-series designs have fallen somewhat out of favor (Borckardt et al., 2008). Once the mainstay of behavior therapists in the 1970s and early 1980s, single-case designs focus on the experimental analysis of behavior (Hersen & Barlow, 1976). Using single-case interventions may provide the establishment of a model of individualized EBP in which the goal would be less the use of scientifically established treatments and more the scientific use of treatment (Gambrill, 2006), thus returning to the roots of behavior therapy and also bridging the gap between research and practice. The APA Division 12 task force includes the use of systematic single-case intervention as one manner from which to glean scientific evidence (Chambless & Hollon, 1998).

Single-case designs allow for multiple observations before and after treatment to provide evidence of patient change and can be accomplished in both clinical settings and research settings (see Borckardt et al., 2008). Single-case studies have natural appeal to practitioners as they can provide information relevant to each client and allow for the comparison of interventions to determine which works best for this client under specific circumstances (Stewart & Chambless, 2010). Additionally, single-case time-series designs can include important manipulations (e.g., randomization) to help ensure a degree of methodological rigor from which researchers can generate causal inferences (Kratochwill & Levin, 2010; Lewin, Lall, & Kratochwill, 2011).

QUALITATIVE METHODS

Qualitative methods offer a window into the complex processes occurring within DI research studies in a manner that purely quantitative studies are unable to provide. Qualitative research "provides a vivid, dense, and full description in the natural language of the phenomenon under study" (Hill, Thompson, & Williams, 1997, p. 518). Rather than identifying *a priori* hypotheses, relationships between phenomena are identified as part of the process of qualitative research. A qualitative approach allows for the "change over time" investigation of DI efforts (Meyer, 2004).

Qualitative methods can be used to "explore and obtain depth of understanding as to the reasons for success or failure to implement evidence-based practice or to identify strategies for facilitating implementation while quantitative methods are used to test and confirm hypotheses based on an existing conceptual model and obtain breadth of understanding of predictors of successful implementation" (Palinkas et al., 2011, p. 44). In this way, qualitative methodology can be used to augment traditional quantitative methods by providing more nuanced contextual information on barriers and/or facilitators. Numerous examples of exemplary use of qualitative methodology exist within DI literature. For example, one study used an ethnographic approach to understand intentions of community clinicians to use EBP (Palinkas et al., 2008). In this study, participant observation and semistructured interviews were used to understand treatment implementation in an effectiveness trial of EBP for depression, anxiety, and conduct problems in youth. Three patterns emerged with regard to participant intention to use EBP: application of treatment with fidelity, abandonment of treatment, and selective application of treatment. Factors associated with these intentions were also explored.

Qualitative research methods, like all methodologies, are not without limitations. Despite increasing attention to the value of such methods, weaknesses include less scientific rigor than quantitative methods and concerns about reliability and validity, analytic techniques used, and quality of produced knowledge (Fitzpatrick & Bolton, 1996; Mays & Pope, 2000).

SUMMARY OF DESIGNS

Each design can be useful when attempting to answer questions relevant to DI processes, and careful consideration of the research question and balancing the strengths and limitations of each design is necessary. A recent review describing elements in studies of EBP implementation in child welfare and mental health settings found RCTs to be the dominant paradigm, with some utilization of mixed methodology. Little use of emerging alternative designs (e.g., PCTs, SMART design) was identified (Landsverk et al., 2010), suggesting that future studies should consider these alternatives.

In a developing area such as DI, researchers might recognize the strengths of established methods but also consider the use of multiple-method research to produce converging results. For example, we agree with Dattilio, Edwards, and Fishman (2010) that each DI study should include (a) an RCT, (b) a qualitative evaluation of the implementation of the study with an emphasis on organizational characteristics, and (c) systematic case studies. Taking a mixed-method approach to DI processes moves the field toward a rapprochement between research and practice (Dattilio et al., 2010). A review of 22 studies utilizing mixed methods in child mental health services research found that mixed methods were used for one of five reasons: (a) to measure intervention and process, (b) to conduct exploratory and confirmatory research, (c) to examine intervention content and context, (d) to understand perspectives of consumers (i.e., practitioners and clients), and (e) to compensate for one set of methods with another (Palinkas et al., 2011). The authors state, "it is the combining of these methods through mixed method designs that is likely to hold the greatest promise for advancing our understanding of why evidence-based practices are not being used, what can be done to get them into routine use, and how to accelerate the improvement of systems of care and practice" (Palinkas et al., 2011).

Outcomes Relevant to Dissemination and Implementation

A number of variables have been examined as both predictors and outcomes within the DI literature and include individual provider (e.g., knowledge, attitudes), organizational (e.g., climate, support), and client variables (e.g., treatment outcome). However, given the present emphasis on DI methods, we focus on reviewing implementation outcomes.

Proctor and colleagues (2011) recommend that DI research focus on implementation outcomes that are conceptually different from service or client outcomes. Specifically, the authors "define implementation outcomes as the effects of deliberate and purposive actions to implement new treatments, practices, and services" (Proctor et al., 2011, p. 65). An emphasis on implementation outcomes is necessary given that such outcomes are indicators of implementation success, are proximal indicators of implementation processes, and are related to service and clinical outcomes (Proctor et al., 2011). Distinguishing between implementation and intervention effectiveness is crucial in DI studies to understand what occurs following implementation (i.e., is failure due to a poorly designed or inappropriate intervention or to an effective practice implemented inadequately). Proctor and colleagues (2011) suggested that there are eight crucial

outcomes to understand the effects of DI studies: acceptability, adoption, appropriateness, feasibility, fidelity, implementation cost, penetration, and sustainability. We suggest adaptation of intervention as an additional outcome of interest.

Acceptability refers to the belief among stakeholders that a particular EBP is acceptable and up to standards. Proctor and colleagues (2011) distinguish acceptability from satisfaction, stating that acceptability is more specific to a particular set of practices. Additionally, acceptability is fluid in that it changes with experience (e.g., before to after implementation). Acceptability can be measured at the individual provider, organizational, and client levels. One example of an instrument that measures this construct includes the Evidence-Based Practice Attitude Scale (EBPAS; Aarons, 2004).

Adoption refers here to "the intention, initial decision, or action to try or employ an innovation or EBP" (Proctor et al., 2011). Adoption is measured at the individual or organizational level and refers here to the same construct as delineated in the RE-AIM model. Standardized measures of adoption have yet to be identified, and time criteria have not been specified (i.e., when does adoption become routine practice).

Appropriateness refers to the compatibility of an EBP for a given setting, provider, or consumer. The constructs of appropriateness and acceptability overlap but are also distinct given that an EBP can be appropriate but not acceptable (Proctor et al., 2011). Standardized measures of appropriateness have not been identified.

Feasibility refers to the extent to which an EBP can be used effectively within a service system (Proctor et al., 2011) and can be assessed from an individual and organizational level. Measures of feasibility have not been identified.

Implementation cost refers to the cost of an implementation effort and varies with regard to delivery, complexity of the EBP, and particular service setting. The few studies that have reported on implementation cost have quantified cost by intervention component. However, direct measures of implementation cost are not currently widely used (Proctor et al., 2011). One possible strategy to be used is the English cost calculator, a method used to calculate the cost of core work activities and administrative costs, in order to inform administrators when making implementation decisions (Chamberlain et al., 2011). It is likely that the DI field can benefit from work in health economics to advance this area.

Penetration refers to "the integration of a practice within a service setting and its subsystems" (Proctor et al., 2011)—in other words, how widely used a particular practice is within an organization, conceptually similar to the "reach" component in the RE-AIM framework. Direct measures of penetration have not been identified.

Any successful DI effort should result not only in the EBP being implemented within the community, but also *sustainability* over time if found to be effective. This construct is akin to the "maintenance" component in the RE-AIM model and is directly addressed in the PRISM model.

It is likely that sustained programs are better situated to yield sustained effects. Sustainability is also crucial because outcomes may not realistically be achieved or detected within the timeframe permitted by traditional research studies or the grants that typically support them, particularly if the intervention targets behavioral change or community-level mental health outcomes (Pluye, Potvin, & Denis, 2004). Moreover, the recurrent discontinuation of promising or effective programs can have deleterious consequences for a community, specifically with regard to willingness to support future projects (Pluye et al., 2004; Shediac-Rizkallah & Bone, 1998).

Sustainability has been operationalized multiple ways, including the continuation of program activities, the maintenance of intended benefits for the target population, and the development of community capacity (Scheirer, 2005; Shediac-Rizkallah & Bone, 1998). Altman (1995, p. 527) has proposed an especially clear definition:

> Sustainability is … defined as the infrastructure that remains in a community after a research project ends. Sustainability includes consideration of *interventions* that are maintained, *organizations* that modify their actions as a result of participating in research, and *individuals* who, through the research process, gain knowledge and skills that are used in other life domains.

This conceptualization highlights the relationship between a program and the setting in which it is implemented and emphasizes that systemic change at multiple levels ought to be a goal of any intervention. Thus, thinking about sustainability ought to reflect enduring change at the community level, as should the ways in which sustainability is planned for and measured.

Too often, DI research is viewed as a linear process, culminating in the sustainability phase. More

effective, however, is to view sustainability as a process that unfolds alongside the research effort (Pluye et al., 2004). From this perspective, planning for sustainability becomes part of planning for the DI process more generally (Adelman & Taylor, 2003; Altman, 1995; Pluye et al., 2004), and this planning is best informed by an understanding of factors believed to influence sustainability, among them (a) the presence of a program "champion" or change agent (Adelman & Taylor, 2003; Scheirer, 2005; Shediac-Rizkallah & Bone, 1998), (b) the extent to which the program is compatible with an organization's values or mission (Scheirer, 2005), (c) the extent to which the program is integrated into the structures and routines of an organization or community (Adelman & Taylor, 2003; Shediac-Rizkallah & Bone, 1998), (d) the extent to which community members perceive the program as beneficial and support it (Altman, 1995; Scheirer, 2005), and (e) flexibility to modify the program over time (Scheirer, 2005).

All but the last of these factors can benefit from an ongoing collaboration between researcher and community: "The literature overwhelmingly shows a positive relationship between community participation and sustainability" (Shediac-Rizkallah & Bone, 1998, p. 103). Early involvement of community members in the research process can help researchers appreciate the needs of the community, thereby enabling them to study and develop interventions that better meet those needs (Altman, 1995; Durlak & DuPre, 2008). This, in turn, can increase the willingness among community members and groups to take ownership of the intervention and sustain it beyond the initial funding period (Altman, 1995). To date, measures of sustainability are not available.

Fidelity refers to the implementation of an EBP as specified by treatment developers. Measuring provider adherence and competence/skill has become standard procedure to determine treatment fidelity (Kendall & Comer, 2011; Perpepletchikova & Kazdin, 2005). Adherence refers to the degree to which a clinician follows the procedures of an EBP, whereas competence refers to the level of skill demonstrated by the clinician in the delivery of treatment (Perepletchikova & Kazdin, 2005). Adherence and competence are typically measured by independent evaluators based on in-session clinician behavior. Illustrative examples of fidelity measures include the Cognitive Therapy Scale (Young & Beck, 1980) and the Motivational Interviewing Treatment Integrity scale (Moyers, Martin, Catley, Harris, & Ahluwalia, 2003). One difficulty with measuring fidelity includes varying fidelity measures across treatment modality.

The emphasis on fidelity has come under criticism. A recent meta-analysis suggests that neither adherence nor competence is significantly related to patient outcomes (Webb, DeRubeis, & Barber, 2010). Possible explanations of this puzzling finding include limited variability on adherence and competence ratings within RCTs included in this meta-analysis (therapists are trained to criterion and monitored, resulting in a limited range) and the possibility of a curvilinear relationship between fidelity and outcomes. However, much is unknown about the causal role of specific treatment interventions on specific outcomes (Morgenstern & McKay, 2007), and more dismantling studies are needed to understand the relative contribution of various therapeutic procedures on outcomes. Given the current literature, it is premature to conclude that fidelity to EBP is unimportant in DI efforts, but further empirical study is necessary.

The question of *adaptation* of treatments to particular settings has been raised with regard to fidelity. Adaptation has been defined as intentional or unintentional additions, deletions, or modifications of a program (Center for Substance Abuse Prevention, 2002). The term "re-invention" has been used (Rogers, 1995), often interchangeably. Most researchers agree that adaptation is not inherently negative; it is often beneficial to make certain changes to better address the needs, culture, and context of the local environment (Bauman, Stein & Ireys, 1991; Castro, Barrera, & Martinez, 2004; Center for Substance Abuse Prevention, 2002; Ozer, Wanis & Bazell, 2010; Rogers, 1995). In fact, there is evidence to suggest that adaptation can serve to increase both the effectiveness of an intervention (e.g., McGraw, Sellers, Stone & Bebchuk, 1996) and the likelihood that an intervention is sustained over time (e.g., Scheirer, 2005), which may be a consequence of increasing the relevance of the intervention for the target population (Castro et al., 2004; Ozer et al., 2010).

When we shift our attention to the process by which individuals and organizations implement EBP, a key issue that arises is the extent to which the programs or practices being used in fact resemble those upon which the evidence was based. Despite findings from a recent meta-analysis (Webb et al., 2010), a number of studies have demonstrated that a high level of fidelity to an intervention's design has been linked to improved outcomes (Battistich,

Schaps, Watson, & Solomon, 1996; Blakely et al., 1987; Botvin, Baker, Dusenbury, Tortu, & Botvin, 1990; Durlak & DuPre, 2008; Rohrbach, Graham, & Hansen, 1993), and there are those who insist that absolute fidelity must be maintained (O'Connor, Small, & Cooney, 2007). Many researchers acknowledge that what matters most is fidelity to an intervention's core components or causal mechanism(s). In other words, testing interventions in real-world settings requires a balancing act, of sorts, between preserving an intervention's core components and making needed adaptations given the local context (i.e., flexibility within fidelity; Bauman et al., 1991; Center for Substance Abuse Prevention, 2002; Green & Glasgow, 2006; Kendall & Beidas, 2007; Kendall, Gosch, Furr, & Sood, 2008).

Rogers (1995) noted that some amount of re-invention is inevitable among adopters of innovations; for example, several studies report that adaptations are the norm when implementing school-based interventions (Datnow & Castellano, 2000; Dusenbury, Brannigan, Hansen, Walsh, & Falco, 2005; Larsen & Samdal, 2007; Ozer et al., 2010; Ringwalt, Ennett, Vincus, & Simons-Rudolph, 2004). That said, the true prevalence of adaptations is unknown because they are not reported consistently. Durlak and DuPre (2008) found that only 3 of 59 studies assessing the impact of implementation on intervention outcomes reported on adaptation, whereas 37 reported on fidelity.

In light of this, those involved in DI research must take care to document the adaptation process. According to the Center for Substance Abuse Prevention (2002), the following steps have been proposed to guide the process of adapting programs to new settings: (a) identify the theory of change underlying the program, (b) identify the components that are essential to the program (i.e., its "core" components), (c) identify appropriate adaptations given the local circumstances, (d) consult with the program developer regarding the previous steps, (e) consult with local stakeholders, and (f) develop a plan for implementation, including a plan for assessing the fidelity/adaptation balance.

The task is not without challenges. First, few interventions adequately delineate which components are *core* (Durlak & DuPre, 2008), making it difficult to determine whether a proposed adaptation may threaten the very mechanism that makes the intervention work. Those involved in DI research are urged to work in tandem with program developers, requesting, if necessary, that they conduct some manner of core component analysis. Ideally, program developers would not only identify those elements central to the program's theory of change that must remain intact, but also articulate the range of acceptable adaptations (Green & Glasgow, 2006). Second, the definition provided earlier—which encompasses additions, deletions, and modifications to the program model—may lead to some confusion regarding what actually *counts* as an adaptation. For instance, how should we distinguish between an *addition* to a program's model and a separate but related practice taking place alongside the program, within the same organization?

These challenges demand a thoughtful and deliberate implementation process, in which researchers work closely with local stakeholders to plan for the implementation of EBP. During this process, consideration should be given to both the local conditions that make adaptations appropriate in practice, as well as the extent to which they may be permissible by the theory underlying the intervention (Green & Glasgow, 2006). Finally, descriptions and rationales for adaptations must be documented so that implementation can be more meaningfully evaluated and outcomes can be interpreted more accurately.

Measures

Variables of interest in DI research vary from those in other related areas. Accordingly, measures explicit to DI research have emerged and made it possible to measure constructs from an ecological perspective including provider, client, and organizational variables (Table 5.2). Measures specific to DI processes (as in Proctor et al., 2011) also exist. For further discussion of DI measures, see Lewis, Comtois, and Krimer (2011).[2]

MEASURES AT THE PROVIDER LEVEL

Provider Attitudes

Measure of Disseminability (MOD; Trent, Buchanan, & Young, 2010), a 32-item self-report measure, assesses therapists' attitudes toward the adoption of a particular EBP on a scale from 1 (not at all) to 7 (very much). The MOD is based upon a three-factor model (treatment evaluation, level of comfort, and negative expectations) that has been studied using exploratory and confirmatory factor analysis (Trent et al., 2010). Psychometric properties include strong retest reliability (.93) and internal consistency (.73 to .83; Trent et al., 2010).

Evidence-Based Practice Attitude Scale (EBPAS; Aarons, 2004), a 15-item self-report measure, assesses therapists' attitudes toward the adoption and implementation of EBP on a scale from 0 (not at all)

Table 5.2 A Comparison of DI Measures

	Measure Features									
Measure	Provider Level	Provider Attitudes	Provider Knowledge	Provider Fidelity	Organizational Level	Implementation Process	Client Level	Psychometrically Investigated	Freely Available	Website
MOD	■	■						■	■	
EBPAS	■	■						■	■	
MPAS	■	■						■	■	
ASA	■	■						■	■	
TX-CHAT	■	■						■	■	
KEBSQ	■		■					■	■	http://www.childfirst.ucla.edu/resources.html
CBT-KQ	■		■							
TPOCS-S	■			■				■	■	
ORC					■			■	■	http://www.ibr.tcu.edu/evidence/evi-orc.html
OSC	■				■			■	■	
ORCA	■				■	■		■	■	http://www.implementationscience.com/content/4/1/38
AII	■	■				■		■	■	
SHAY					■	■		■	■	
TCAT					■	■			■	http://www.ibr.tcu.edu/pubs/datacoll/commtrt.html
OS							■	■	■	http://www.mh.state.oh.us/what-we-do/protect-and-monitor/consumer-outcomes/instruments/index.shtml
CIS							■	■	■	http://www.dhs.state.il.us/page.aspx?item=32589
PROMIS							■	■	■	http://www.nihpromis.org/

Note: MOD = Measure of Disseminability (Trent, Buchanan, & Young, 2010); EBPAS = Evidence Based Practice Attitude Scale (Aarons, 2004); MPAS = Modified Practitioner Attitude Scale (Chorpita et al., 2004); ASA = Attitudes Toward Standardized Assessment Scales (Jensen-Doss & Hawley, 2010); TX-CHAT = Texas Survey of Provider Characteristics and Attitudes (Jensen-Doss, Hawley, Lopez, & Osterberg, 2009); KEBSQ = Knowledge of Evidence Based Services Questionnaire (Stumpf et al., 2009); CBT-KQ = Cognitive-Behavioral Therapy Knowledge Quiz (Latham, Myles, & Ricketts, 2003; Myles, Latham, & Ricketts, 2003); TPOCS-S = Therapy Process Observational Coding System for Child Psychotherapy Strategies Scale (McLeod, 2001); ORC = Organizational Readiness for Change (Institute for Behavioral Research, 2002); OSC = Organizational Social Context (Glisson et al., 2008); ORCA = Organizational Readiness to Change Assessment (Helfrich, Li, Sharp, & Sales, 2009); AII = Adopting Innovation Instrument (Moore & Benbasat, 1991); SHAY = State Health Authority Yardstick (Finnerty et al., 2009); TCAT = Treatment Cost Analysis Tool (Flynn et al., 2009); OS = Ohio Scales (Ogles, Lunnen, Gillespie, & Trout, 1996); CIS = Columbia Impairment Scale (Hurt, Arnold & Aman, 2003); PROMIS = Patient-Reported Outcomes Measurement Information System (Cella et al., 2010).

■ = Feature characterizes model

to 4 (to a great extent). The EBPAS maps onto four subscales: appeal, requirements, openness, and divergence (Aarons, 2004). Appeal refers to the extent to which a therapist would adopt a new practice if it is intuitively appealing. Requirements refers to the extent to which a therapist would adopt a new practice if required by his or her organization or legally mandated. Openness is the extent to which a therapist is generally receptive to using new interventions. Divergence is the extent to which a therapist perceives research-based treatments as not useful clinically (Aarons, 2004). The EBPAS demonstrates good internal consistency (Aarons, 2004), subscale alphas range from .59 to .90 (Aarons & Sawitzky, 2006), and its validity is supported by its relationship with both therapist-level attributes and organizational characteristics (Aarons, 2004). Recently, a 50-item version of the EBPAS (EBPAS-50) has been developed and includes an additional eight factors: limitations, fit, monitoring, balance, burden, job security, organizational support, and feedback. Exploratory analyses demonstrated high internal consistency among factors (.77 to .92; Aarons, Cafri, Lugo, & Sawitzky, 2010).

Modified Practitioner Attitude Scale (MPAS; Chorpita et al., unpublished measure, 2004) is an eight-item measure created for administration to direct service providers to understand therapists' attitudes toward EBP. Items are measured on a scale from 0 (not at all) to 4 (to a great extent). Items on the MPAS are similar to items on the EBPAS but are specifically worded to avoid references to treatment manuals (e.g., referring to treatments rather than treatment manuals). Psychometric properties for the MPAS suggest adequate internal consistency (.80) and moderate relationship with the EBPAS ($r = .36$). The wording in the MPAS (i.e., not referring to treatment manuals but referring to EBP) may result in differential results in reported provider attitudes (Borntrager, Chorpita, Higa-McMillan, & Weisz, 2009).

Attitudes Toward Standardized Assessment Scales (ASA; Jensen-Doss & Hawley, 2010) is a 22-item measure created for administration to direct service providers to understand therapists' attitude towards standardized assessment measures often utilized in EBP. Items are measured on a scale from 1 (strongly agree) to 5 (strongly disagree). Items address three factors: benefit over clinical judgment, psychometric quality, and practicality. Benefit over clinical judgment refers to items assessing the extent to which standardized measures provide extra information above and beyond clinical judgment by itself. Psychometric quality refers to clinicians' beliefs about the reliability and validity of standardized measures. Practicality refers to clinicians' belief about the feasibility of using standardized measure in clinical practice. In an initial psychometric evaluation, internal consistency ranged from .72 to .75, and scale structure was corroborated by a confirmatory factor analysis suggesting adequate model fit (RMSEA = .045, CFI = .935). The measure was also found to be predictive of intentions to use evidence-based assessment (Jensen-Doss & Hawley, 2010).

Texas Survey of Provider Characteristics and Attitudes (TX-CHAT; Jensen-Doss, Hawley, Lopez, & Osterberg, 2009) is a 27-item measure created for administration to direct service providers to understand therapists' attitudes toward EBPs that they are currently using in their clinical practice. Items are measured on a scale from 1 (not at all true for me) to 5 (very true for me). Items map onto five subscales: provider's attitudes toward evidence-based treatments, colleagues' attitudes toward evidence-based treatments, agency support for implementation, barriers to implementation, and quality of training. The measure has held up to initial psychometric investigation with adequate alpha's at .69 or above (Jensen-Doss et al., 2009; Lopez, Osterberg, Jensen-Doss, & Rae, 2011).

Provider Knowledge

Knowledge of Evidence Based Services Questionnaire (KEBSQ; Stumpf, Higa-McMillan, & Chorpita, 2009) is a 40-item self-report measure administered to direct service providers to measure their knowledge of EBP. Items on the KEBSQ include practice elements of EBP and non-EBP used in the treatment of four childhood difficulties: (a) anxious/avoidant, (b) depressed/withdrawn, (c) disruptive behavior, and (d) attention/hyperactivity. In this measure, 40 practice elements are listed and practitioners are to classify if a particular practice element (e.g., relaxation) is used in EBP for each of the four difficulties. Each item is scored on a scale from 0 to 4 with a total possible score of 160; higher scores indicate more knowledge of EBP. The measure has acceptable temporal stability (.56), sensitivity to training, and discriminative validity (Stumpf et al., 2009). Overall internal consistency is low (.46; Okamura, Nakamura, McMillan, Mueller, & Hayashi, 2010), but the original authors caution against measuring internal consistency given that each item represents a unique and independent technique that is not necessarily related to other items (Stumpf et al., 2009).

Cognitive-Behavioral Therapy Knowledge Quiz (CBT-KQ; Latham, Myles, & Ricketts, 2003; Myles, Latham, & Ricketts, 2003) is a 26-item self-report multiple-choice measure administered to direct service providers to measure knowledge of CBT in adult patients. Items on the CBT-KQ map onto the following categories: (a) general CBT issues, (b) underpinnings of behavioral approaches, (c) underpinnings of cognitive approaches, (d) practice of behavioral psychotherapy, and (e) practice of cognitive therapy. Each item is scored as correct or incorrect with a total possible score of 26; higher scores indicate more knowledge of CBT. Psychometrics are not yet available.

Provider Intervention Fidelity

Several instruments exist to measure fidelity to specific treatment modalities. For example, for motivational interviewing, one can use the Motivational Interviewing Skill Coding (MISC; Moyers et al., 2003) whereas for cognitive therapy, one can use the Cognitive Therapy Scale (CTS; Young & Beck, 1980), the Cognitive Therapy Scale-Revised (CTS-R; James, Blackburn, & Reichelt, 2001), or the Collaborative Study Psychotherapy Ratings Scale (CSPRS; Hollon et al., 1988). Often, investigators create intervention-specific fidelity measures for the specific EBP they are researching and disseminating (Beidas & Kendall, 2010). Recommendations have been made for using standardized fidelity measures across EBPs; however, there is currently no measure that can be used across EBPs, and as can be seen, often multiple measures exist for the same treatment modality. However, one observational coding system that cuts across modalities for child psychotherapy strategies has been psychometrically explored and is described below.

Therapy Process Observational Coding System for Child Psychotherapy Strategies Scale (TPOCS-S; McLeod, 2001) is a 31-item coding measure intended to allow for description of provision of mental health treatment in practice settings. TPOCS-S subscales differentiate between intervention strategies and include cognitive, behavioral, psychodynamic, family, and client-centered techniques. The TPOCS-S scoring involves "extensiveness ratings of therapeutic interventions designed to measure the degree to which therapists use specific therapeutic interventions during a therapy session" (McLeod & Weisz, 2010, p. 438). Coders observe sessions and indicate the degree to which a therapist engages in each strategy during the whole session from 1 (not at all) to 7 (extensively). Extensiveness ratings include thoroughness and frequency. Thoroughness refers to depth of provision of intervention; frequency refers to how often a therapist provides the intervention during a session. The TPOCS-S has been psychometrically investigated. The measure has shown good interrater reliability (.66 to .95), internally consistent subscales (.74 to .86), and adequate construct validity (McLeod & Weisz, 2010). The TPOCS-S has been used successfully in studies characterizing usual care (Garland et al., 2010).

MEASURES AT THE ORGANIZATIONAL LEVEL

Organizational Readiness for Change (ORC; Institute for Behavioral Research, 2002) is a 129-item instrument that measures organizational characteristics and is gathered through administration to various individuals in an organization. Responses are provided based on a 5-point Likert rating scale ranging from 1 (strongly disagree) to 5 (strongly agree). The 18 scales represent three major domains: motivation, resources, and organizational factors. Motivational factors include program needs, training needs, and pressure for change. Resources include office facilities, staffing, training, equipment, and availability of Internet. Organizational factors include staff attributes and organizational climate. Staff attributes include growth, efficacy, influence, and adaptability; organizational climate includes mission, cohesion, autonomy, communication, stress, and flexibility for change.

Psychometrically speaking, the instrument has shown moderate to high coefficient alphas (range: .56 to .92), and support for the factors has been gleaned from principal component analysis (Lehmen, Greener, & Simpson, 2002). This measure has multiple forms to be administered to various individuals within an organization, such as front-line staff and supervisors. Additionally, the measure has been modified for use in settings other than community mental health centers (e.g., criminal justice). Score profiles can be mapped onto norms, allowing for direct comparisons to other national organizations. Ideally, the measure is administered to at least five individuals in an organization (TCU IBR, 2002).

Organizational Social Context (OSC; Glisson et al., 2008) is a measurement system that quantitatively evaluates the social context of mental health and social services organizations through administration to direct service providers. Specifically, the OSC measures both individual-level (work attitudes, work behavior) and organizational-level (culture) variables, as well as individual and shared

perceptions (climate). Assessing the social context of an organization makes it possible to capture features that may influence service and treatment, clinician morale, and adoption and implementation of EBP.

The OSC has 105 items that form 16 first-order scales and 7 second-order scales. Factors are grouped by structure, culture, psychological and organizational climate, and work attitudes. Culture refers to the norms and values of an organization; climate refers to the impact of a work context on an individual. Work attitudes refer to morale of an individual worker. The measurement of these factors together allows for an understanding of an organization's context and can be compared with norms of national service settings. Confirmatory factor analysis supported these factors; alpha coefficients for scales range from .71 to .94 (Glisson et al., 2008). It is preferable that four or more individuals from an organization complete this assessment for adequate measurement of organizational climate (P. Green, personal communication).

MEASURES SPECIFIC TO DI PROCESSES

The instruments below measure specific constructs relevant to DI processes and either map onto relevant DI models (e.g., PARiHS, DOI) or provide information specific to Proctor and colleagues' (2011) suggested implementation outcomes.

Organizational Readiness to Change Assessment (ORCA; Helfrich, Li, Sharp, & Sales, 2009) operationalizes the core constructs of the PARiHS framework. The ORCA is a 77-item measure that is administered to staff involved in quality improvement initiatives; responses range from 1 (very weak) to 5 (very strong). Items map onto three scales that make up the core elements of the PARiHS framework: (a) strength and extent of evidence, (b) organizational climate, and (c) capacity for internal facilitation of QI program. A three-factor solution was identified via exploratory factor analysis, and reliability (.74 to .95) was acceptable, but further validation is necessary (Helfrich et al., 2009). A follow-up study found the preimplementation ORCA scores to be predictive of low and high implementation rates across sites (Hagedorn & Heideman, 2010).

Adopting Innovation Instrument (Moore & Benbasat, 1991) is a 38-item self-report measure that assesses perceptions a provider may have toward adopting an innovation. In the rigorous development of this instrument, the authors specifically aimed to measure the constructs that Rogers (2004) proposed. Specifically, this instrument contains eight factors: relative advantage, compatibility, ease of use, result demonstrability, image, visibility, trialability, and voluntariness. Psychometrics are adequate with regard to reliability (.71 to .95) and validity, with a principal component analysis identifying seven factors (Moore & Benbasat, 1991).

State Health Authority Yardstick (SHAY; Finnerty et al., 2009) is a 15-item agency-specific behaviorally anchored instrument that assesses systems-level considerations that are relevant to the implementation of EBP. Specifically, the SHAY assesses seven domains: planning, financing, training, leadership, policies and regulations, quality improvement, and stakeholders. Items are rated from 1 (little or no implementation) to 5 (full implementation). The SHAY is intended to be administered by two independent raters who interview multiple informants in an organization. The two raters make independent ratings and then create consensus ratings. Initial evidence partially supports construct and criterion validity of the instrument in assessing state-level facilitators of and/or barriers to EBP implementation (Finnerty et al., 2009).

Treatment Cost Analysis Tool (TCAT; Flynn et al., 2009) is a measure created to assist in cost analysis of outpatient substance abuse treatment programs. To generate cost analysis, the TCAT includes information about client volume, counseling, total program costs, overhead costs, and personnel data. The measure is easy to use and is available through an Excel spreadsheet. This measure provides information on cost effectiveness as suggested by Proctor and colleagues (2011).

MEASURES AT THE CLIENT LEVEL

Given the complexity of DI studies, client measures to address client characteristics and client outcomes should be easy to implement and score, freely available so that their use may be sustained following the research project, and specific to the research question. Several large systems have adopted outcome measures that would be appropriate in DI research studies.

For example, Illinois requires the Ohio Scales (Ogles, Lunnen, Gillespie, & Trout, 1996) and Columbia Impairment Scale (Hurt, Arnold & Aman, 2003) for all children funded by Medicaid. The Ohio Scales (Ogles et al., 1996) focus on efficient administration, scoring, and interpretation. There are three parallel forms of the Ohio Scales that can be completed by the youth, caregiver, and service provider. All forms include questions relating to problem severity, functioning, satisfaction, and hopefulness. These scales were developed not

to diagnose youth but to provide an efficient means of tracking outcomes in community agencies. Psychometric properties are solid with adequate test–retest reliability (.65 to .97) and preliminary validity (Ogles et al., 1996). The Columbia Impairment Scale (CIS; Hurt et al., 2003) focuses on impairment of functioning and assesses how well an individual carries out age-appropriate daily activities. The items are scored on a 4-point scale, with a greater score indicating greater impairment. The CIS can be filled out by either a clinician or a caregiver and demonstrates good internal consistency, test–retest reliability, and validity (Hurt at al., 2003).

An exiting initiative sponsored by the National Institutes of Health also has produced efficient and easily accessible outcome measures that can be utilized in DI studies: the Patient-Reported Outcomes Measurement Information System (PROMIS). The goal of this project is "to develop and evaluate, for the clinical research community, a set of publicly available, efficient, and flexible measurements of patient-reported outcomes, including health-related quality of life" (Cella et al., 2010, p. 1180). Content areas for items include physical health (e.g., fatigue), mental health (e.g., anxiety), and social health (e.g., social function). These items are available as paper-and-pencil measures and computer adaptive tests. Large-scale testing of PROMIS items suggests good reliability and validity. A larger discussion of PROMIS is beyond the scope of this chapter, but these tools may be particularly well suited for DI studies given their brevity, ability to be tailored to particular populations, and ease of use. For example, if one is interested in studying the DI of CBT for youth anxiety disorders, one could use the pediatric PROMIS anxiety and depressive symptoms scales to measure outcomes (Irwin et al., 2010).

Conclusion and Future Directions

DI science is a relatively new area of inquiry within mental health services research that strives to understand the key mechanisms and processes needed to expand the utilization of EBP in community mental health settings. DI research aims to bridge the research-to-practice gap that prevents knowledge and practices of effective treatments from reaching many people in need (Weisz, Donenberg, Han, & Kauneckis, 1995). Researchers focus on the need to systematically study the process of DI to increase the use of best practices in community mental health settings (Schoenwald & Hoagwood, 2001; Schoenwald, Hoagwood, Atkins, Evans, & Ringeisen, 2010).

Several models relevant to DI research were described. Comprehensive models (e.g., CFIR; Damschroder et al., 2009) provide heuristics as to areas and levels of study, while more specific models (e.g., TPB; Ajzen, 1988, 1991) describe specific targets that might be vital to understand DI mechanisms. The comprehensive models underscore the importance of considering multiple levels of change and organization, leading to the need for complex studies that address not only *whether* an implemented intervention has the desired effect, but also *how* the context affects the changes (or lack thereof) that may be a result of the intervention. This necessitates the careful and thoughtful assessment of fidelity and a thorough understanding of the issues that are inherent in fidelity measurement (e.g., What are the core elements of the intervention? Is ongoing quality assessment incorporated into the DI process? Can interventions be adapted with fidelity?). With regard to adaptation, empirical questions to tackle include: What adaptations, for whom, result in improved client outcomes? Do adaptations result in higher rates of implementation or sustainability of EBP? When does flexibility in implementation of an EBP become infidelity (Kendall & Beidas, 2007; Kendall et al., 2008)?

The specific models reviewed suggest possible targets of intervention to optimize DI efforts. For example, models emphasizing the organizational context of DI efforts (e.g., ARC; Glisson & Schoenwald, 2005) suggest that key components of the organizational context such as norms and expectations within the setting may influence DI outcomes. An organizational perspective includes how to influence and support an organization in the adoption and implementation of EBP. Traditionally, this perspective has included an understanding of the structures needed to support new models and learning and infrastructures that can support new models of mental health services. For example, providing facilitation to an organization in the creation of the structures and infrastructures needed to support a new intervention model increased the likelihood that a particular intervention was adopted and more clients improved following implementation (Glisson et al., 2010).

Other important organizational considerations include the role of social networks within DI efforts. Given that adoption of EBP may be a slow process, program response needs to be understood as unfolding over time, requiring longitudinal studies that account for the differential adoption of interventions. In addition, if programs are not adopted

throughout a social system such as an agency or school, this may suggest that the program is not seen by a sufficient number of members as appropriate to their needs. This could lead to a series of questions as to how to adapt programs or how to activate key opinion leaders to influence mental health program use to inform DI efforts throughout a social system (Atkins et al., 2008). Key questions with regard to how organizational context may influence DI outcomes include: How do organizational constructs (e.g., organizational support) operate to facilitate or impede DI? How can knowledge of barriers/facilitators be used to coordinate and augment DI of EBP in community mental health settings? How can organizational interventions effectively improve DI efforts? How can social networks be used to augment DI?

A fundamental issue that arises when taking an organizational perspective is the natural tension between the adaptability of a services setting and the adaptability of a new intervention. There is often an implicit assumption that a service setting is ready to adopt a new intervention. However, if one takes an ecological perspective, there is an active transactional interplay between an organization and a new intervention, with the organization influencing the intervention and the intervention influencing the organization. For example, the organization is likely to be constrained by the structure of the agency, staffing, and budget issues, whereas intervention delivery may be constrained by the common elements that are required to effect change. How and what changes at each level is an empirical question that can enhance the understanding of DI processes and mechanisms. Research that addresses and resolves this tension is paramount.

As stated earlier, the added complexity of including multiple levels of change (i.e., individual, organizational) within a study calls for research methods and design that may stray from the traditional models or "gold standard" of RCTs. Although it remains important to assess and evaluate client outcomes, there are several methods to augment traditional RCT designs, as well as alternative designs (e.g., PCTs). Research on the development of DI-specific methods is sorely needed. Choosing a specific research design requires consideration of the most effective method and design to answer the specific research questions, the strengths and weaknesses of each design, the context of the research, and the available resources. Relying on mixed-method designs may be optimal given that different levels of inquiry may address various questions within the same research study. Finally, there are both proximal and distal outcomes that are relevant for DI research. For example, measuring organizational change or therapist attitude change is a proximal outcome, whereas improved client outcome is the distal outcome.

Despite the many challenges, DI research has an important place in the field of mental health services research. The primary goal of DI research is to identify key processes and mechanisms for change while spreading EBP into community mental health settings (i.e., dissemination), resulting in uptake and adoption of such new technologies (i.e., implementation) that is sustainable (i.e., is maintained). The public policy implications of such empirical inquiry are substantial, given the unmet mental health needs within the U.S. population. One study found that only 21% of in-need children received mental health services within a year and that uninsured youth were especially vulnerable (Kataoka, Zhang, & Wells, 2002).

The public policy implications of DI research suggest that policymakers can, and will, play a key role in shaping the future of DI efforts. Recently, policymakers have been moving from passive users of technology to active participants in a process of DI. For example, Illinois policymakers are insisting that mental health providers implement EBP (e.g., Illinois requires that agencies receiving grants to implement school-based mental health services utilize EBP). This results in the formation of a critical relationship between policy and DI efforts and also provides an opportunity for research and policy to inform one another. Future research can include key questions with regard to public policy such as: How can researchers engage and capitalize on the push policymakers are currently making for the use of EBP? How can researchers partner with policymakers to ensure that efforts are indeed effective and sustainable?

New knowledge is a key feature of all research. DI research can contribute new knowledge both through an understanding of the support and monitoring structures that are needed to support DI of effective practices and the natural processes that support DI, such as social networks and key opinion leaders. Mental health service settings can be transformed with potentially enormous impact on the public health of the general population.

Notes

1. The term "model" encompasses theories, models, and frameworks in this chapter.

2. We thank Cara Lewis, Katherine Comtois, and Yekaterina Krimer for the guidance they provided in the measures section of the chapter: they and the Seattle Implementation Resource

Conference are preparing a comprehensive repository of measures and were generous in discussing these measures with us.

References

Aarons, G. (2004). Mental health provider attitudes toward adoption of evidence-based practice: The Evidence-Based Practice Attitude Scale (EBPAS). *Mental Health Services Research*, *6*, 61–74. doi:10.1023/B:MHSR.0000024351.12294.65

Aarons, G., Cafri, G., Lugo, L., & Sawitzky, A. (2010). Expanding the domains of attitudes towards evidence-based practice: The Evidence Based Practice Attitude Scale-50. *Administration and Policy in Mental Health and Mental Health Services, 39,* 331–340. doi: 10.1007/s10488-010-0302-3

Aarons, G., & Sawitzky, A. C. (2006). Organizational culture and climate and mental health provider attitudes toward evidence-based practice. *Psychological Services*, *3*, 61–72. doi: 10.1037/1541–1559.3.1.61

Adelman, H. S., & Taylor, L. (2003). On sustainability of project innovations as systemic change. *Journal of Educational & Psychological Consultation*, *14*, 1–25. doi:10.1207/S1532768XJEPC1401_01

Ajzen, I. (1988). *Attitudes, personality and behavior*. Milton Keynes, England, Open University Press: Chicago, Dorsey Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. doi:10.1016/0749-5978(91)90020-T

Altman, D. G. (1995). Sustaining interventions in community systems: On the relationship between researchers and communities. *Health Psychology*, *14*, 526–536. doi:10.1037/0278-6133.14.6.526

American Psychological Association. (2005). *American Psychological Association policy statement on evidence-based practice in psychology*. Retrieved from http://www.apa.org/practice/resources/evidence/evidence-based-statement.pdf

Armitage, C. J., & Conner, M. (2001). Efficacy of the Theory of Planned Behaviour: a meta-analytic review. *British Journal of Social Psychology*, *40*, 471–499. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11795063

Atkins, M. S., Frazier, S., & Cappella, E. (2006). Hybrid research models. Natural opportunities for examining mental health in context. *Clinical Psychology: Science and Practice*, *13*, 105–107. doi:10.1111/j.1468-2850.2006.00012.x

Atkins, M. S., Frazier, S. L., Leathers, S. J., Graczyk, P. A., Talbott, E., Jakobsons, L.,. . . Bell, C. (2008). Teacher key opinion leaders and mental health consultation in low-income urban schools. *Journal of Consulting and Clinical Psychology*, *76*, 905–908. doi: 10.1037/a0013036

Battistich, V., Schaps, E., Watson, M., & Solomon, D. (1996). Prevention effects of the child development project: Early findings from an ongoing multisite demonstration trial. *Journal of Adolescent Research*, *11*, 12–35. doi:10.1177/0743554896111003

Bauman, L. J., Stein, R. E., & Ireys, H. T. (1991). Reinventing fidelity: The transfer of social technology among settings. *American Journal of Community Psychology*, *19*, 619–639. doi:10.1007/BF00937995

Beidas, R. S., & Kendall, P. C. (2010). Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. *Clinical Psychology: Science and Practice*, *17*, 1–30. doi:10.1111/j.1468-2850.2009.01187.x

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity–adaptation debate: implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253–268. doi:10.1007/BF00922697

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. doi: 10.1037/0003-066X.63.2.77

Borntrager, C. F., Chorpita, B. F., Higa-McMillan, C., & Weisz, J. R. (2009). Provider attitudes toward evidence-based practices: are the concerns with the evidence or with the manuals? *Psychiatric Services*, *60*, 677–681. doi: 10.1176/appi.ps.60.5.677

Botvin, G. J., Baker, E., Dusenbury, L., Tortu, S., & Botvin, E. M. (1990). Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a 3-year study. *Journal of Consulting and Clinical Psychology*, *58*, 437–446. doi:10.1037//0022-006X.58.4.437

Carroll, K., & Rounsaville, B. (2003). Bridging the gap: A hybrid model to link efficacy and effectiveness research in substance abuse treatment. *Psychiatric Services*, *54*, 333–339. doi:10.1176/appi.ps.54.3.333

Casper, E. (2007). The theory of planned behavior applied to continuing education for mental health professionals. *Psychiatric Services*, *58*, 1324–1329. doi:10.1176/appi.ps.58.10.1324

Castro, F. G., Barrera, M., Jr., & Martinez, C. R., Jr. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science*, *5*, 41–45. doi:10.1023/B:PREV.0000013980.12412.cd

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B. Yount, S… PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*, 1179–1194.

Center for Substance Abuse Prevention. (2002 Conference Edition). *Finding the Balance: Program Fidelity and Adaptation in Substance Abuse Prevention. Executive Summary of a State-of-the-Art Review*. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Substance Abuse Prevention.

Chamberlain, P., Snowden, L. R., Padgett, C., Saldana, L., Roles, J., Holmes, L.,… Landsverk, J. (2011). A strategy for assessing costs of implementing new practices in the child welfare system: Adapting the English cost calculator in the United States. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 24–31. doi: 10.1007/s10488-010-0318-8

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9489259.

Chow, S. C., & Chang, M. (2008). Adaptive design methods in clinical trials—a review. *Orphanet Journal of Rare Diseases*, *3*, 11. doi: 10.1186/1750-1172-3-11

Damschroder, L. J. (Feb. 10, 2011). The role and selection of theoretical frameworks in implementation research. Retrieved Feb. 10, 2011. from www.hsrd.research.va.gov/for_researchers/cyber… /eis-021011.pdf

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health

services research findings into practice: a consolidated framework for advancing *implementation science*. Implementation Science, *4*. doi: 10.1186/1748-5908-4-50

Datnow, A., & Castellano, M. (2000). Teachers' responses to success for all: How beliefs, experiences, and adaptations shape implementation. *American Educational Research Journal*, *37*, 775–799. doi:10.2307/1163489

Dattilio, F. M., Edwards, D. J., & Fishman, D. B. (2010). Case studies within a mixed methods paradigm: toward a resolution of the alienation between researcher and practitioner in psychotherapy research. *Psychotherapy*, *47*, 427–441. doi: 10.1037/a0021181

Dimeff, L., Koerner, K., Woodcock, E., Beadnell, B., Brown, M., Skutch, J.,… Harned, M. (2009). Which training method works best? A randomized controlled trial comparing three methods of training clinicians in dialectical behavior therapy skills. *Behavior Research and Therapy*, *47*, 921–930. doi:10.1016/j.brat.2009.07.011

Dingfelder, H. E., & Mandell, D. S. (2010). Bridging the research-to-practice gap in autism intervention: An application of diffusion of innovation theory. *Journal of Autism and Developmental Disorders*, *41*(5), 597–609. doi: 10.1007/s10803-010-1081-0.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350. doi:10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*, 308–313. doi:10.1093/her/cyg134

Feldstein, A., & Glasgow, R. (2008). A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. *Joint Commission Journal on Quality and Patient Safety*, *34*, 228–242.

Finnerty, M., Rapp, C., Bond, G., Lynde, D., Ganju, V., & Goldman, H. (2009). The State Health Authority Yardstick (SHAY). *Community Mental Health Journal*, *45*, 228–236. doi:10.1007/s10597-009-9181-z

Fitzpatrick, R., & Boulton, M. (1996). Qualitative research in healthcare: I. The scope and validity of methods. *Journal of Evaluation in Clinical Practice*, *2*, 123–130.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*, 531–540. doi: 10.1177/1049731509335549

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, The Louis de la Parte Florida Mental Health Institute. Department of Child & Family Studies. Retrieved from http://nirn.fpg.unc.edu/

Flynn, P., Broome, K., Beaston-Blaakman, A., Knight, D., Horgan, C., & Shepard, D. (2009). Treatment Cost Analysis Tool (TCAT) for estimating costs of outpatient treatment services. *Drug and Alcohol Dependence*, *100*, 47–53. doi:10.1016/j.drugalcdep.2008.08.015

Freedman, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine*, *16*, 141–145. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3600702

Gambrill, E. (2006). *Social work practice: A critical thinker's guide* (2nd ed.) New York: Oxford University Press.

Garland, A., Brookman-Frazee, L., Hurlburt, M., Accurso, E., Zoffness, R., Haine-Schlagel, R.,…Ganger, W. (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services*, *61*, 788–795. doi: 10.1176/appi.ps.61.8.788

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health*, *89*, 1322–1327. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1508772&tool=pmcentrez&rendertype=abstract

Glisson, C., Landsverk, J., Schoenwald, S., Kelleher, K., Hoagwood, K., Mayberg, S.,… The Research Network on Youth Mental Health (2008). Assessing the Organizational Social Context (OSC) of mental health services: Implications for research and practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*, 98–113. doi: 10.1007/s10488-007-0148-5

Glisson, C., & Schoenwald, S. K. (2005). The ARC organizational and community intervention strategy for implementing evidence-based children's mental health treatments. *Mental Health Services Research*, *7*, 243–259. doi: 10.1007/s11020-005-7456-1

Glisson, C., Schoenwald, S. K., Hemmelgarn, A., Green, P., Dukes, D., Armstrong, K. S., & Chapman, J. E. (2010). Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. *Journal of Consulting and Clinical Psychology*, *78*, 537–550. doi: 10.1037/a0019160

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Evaluation & the Health Professions*, *29*, 126–153. doi:10.1177/0163278705284445

Green, L. W., Ottoson, J. M., Garcia, C., & Hiatt, R. A. (2009). Diffusion theory, and knowledge dissemination, utilization, and integration in public health. *Annual Review of Public Health, 30,* 151–174. doi: 10.1146/annurev.publhealth.031308.100049

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: systematic review and recommendations. *Milbank Quarterly*, *82*, 581–629. doi: 10.1111/j.0887-378X.2004.00325.x

Grol, R., Bosch, M., Hulscher, M., Eccles, M., & Wensing, M. (2007). Planning and studying improvement in patient care: The use of theoretical perspectives. *Milbank Quarterly*, *85*, 93–138.

Hagedorn, H., & Heideman, P. (2010). The relationship between baseline Organizational Readiness to Change Assessment subscale scores and implementation of hepatitis prevention services in substance use disorders treatment clinics: a case study. *Implementation Science*, *5*. Retrieved from http://www.implementationscience.com/content/5/1/46

Haider, M., & Kreps, G. (2004). Forty years of diffusion of innovations: Utility and value in public health. *Journal of Health Communication*, *9*, 3–11. doi: 10.1080/10810730490271430

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: how "out of control" can a randomised controlled trial be? *British Medical Journal*, *328*, 1561–1563. doi: 10.1136/bmj.328.7455.1561

Helfrich, C. D., Damschroder, L. J., Hagedorn, H. J., Daggett, G. S., Sahay, A., Ritchie, M.,… Stetler, C. B. (2010). A critical synthesis of literature on the Promoting Action on

Research Implementation in Health Services (PARIHS) framework. *Implementation Science*, *82*. doi: 10.1186/1748-5908-5-82

Helfrich, C. D., Li, Y., Sharp, N., & Sales, A. (2009). Organizational readiness to change assessment: Development of an instrument based on the Promoting Action Research in Health Services (PARiHS) framework. *Implementation Science*, *4*. doi: 10.1186/1748-4-38

Hersen, M., & Barlow, D. (1976). *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press.

Hill, C., Thompson, B., & Williams, E. (1997). A guide to conducting consensual qualitative research. *Counseling Psychologist*, *25*, 517–572. doi:10.1177/0011000097254001

Hoagwood, K., Burns, B., & Weisz, J. (2002). A profitable conjunction: From science to service in children's mental health. In B. Burns & K. Hoagwood (Eds.), Community treatment for youth: Evidence-based interventions for severe emotional and behavioral disorders (pp. 328–338). New York: Oxford University Press.

Hollon, S., Evans, D., Auerbach, A., DeRubeis, R., Elkin, I., Lowery, A.,… Piasecki, J. (1988). *Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal therapy, and clinical management pharmacotherapy.* Unpublished manuscript, University of Minnesota, Twin Cities Campus.

Hurt, E., Arnold, L. E., & Aman, M. G. (2003). Clinical instruments and scales in pediatric psychopharmacology. In A. Martin, L. Scahill, & C. Kratochvil, (2nd ed.), *Pediatric psychopharmacology: Principles and practice* (pp. 389–406). New York: Oxford University Press.

Institute of Behavioral Research (2002). Organizational readiness for change. Texas Christian University, Fort Worth, Texas.

Irwin, D., Stucky, B., Langer, M., Thissen, D., DeWitt, E., Jin-Shei, L.,… DeWalt, D. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*, 595–607. doi:10.1007/s11136-010-9619-3

James, I., Blackburn, I., & Reichelt, F. (2001). *Manual for the Revised Cognitive Therapy Scale (CTS-R)* (2nd ed.). Unpublished manuscript, available from Ian James, Centre for the Health of the Elderly, Newcastle General Hospital, Westgate Road, Newcastle NE4 6BE.

Jensen-Doss, A., & Hawley, K. (2010). Understanding barriers to evidence-based assessment: Clinician's attitudes toward standardized assessment tools. *Journal of Clinical Child & Adolescent Psychology*, *39*, 885–896. doi:10.1080/15374416.2010.517169

Jensen-Doss, A., Hawley, K., Lopez, M., & Osterberg, L. (2009). Using evidence-based treatments: The experiences of youth providers working under a mandate. *Professional Psychology: Research and Practice*, *40*, 417–424. doi:10.1037/a0014690

Kataoka, S., Zhang, L., & Wells, K. (2002). Unmet need for mental health care among US children: Variation by ethnicity and insurance status. *American Journal of Psychiatry*, *159*, 1548–1555. doi: 10.1176/appi.ajp.159.9.1548

Kauth, M., Sullivan, G., Blevins, D., Cully, J., Landes, R., Said, Q., & Teasdale, T. (2010). Employing external facilitation to implement cognitive behavioral therapy in VA clinics: a pilot study. *Implementation Science*, *5*. Retrieved from http://www.implementationscience.com/content/5/1/75

Kendall, P. C., & Beidas, R. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology: Research and Practice*, *38*, 13–20.

Kendall, P. C., & Chambless, D. (Eds.) (1998). Empirically supported psychological therapies, *Journal of Consulting and Clinical Psychology*, *66*, entire issue.

Kendall, P. C., & Comer, J. S. (2011). Research methods in clinical psychology. In D. Barlow (Ed.), Oxford handbook of clinical psychology (pp. 52–75). New York: Oxford University Press.

Kendall, P. C., Gosch, E., Furr, J., & Sood, E. (2008). Flexibility within fidelity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*, 987–993.

Kitson, A., Harvey, G., & McCormack, B. (1998). Enabling the implementation of evidence based practice: a conceptual framework. *Quality in Health Care*, *7*, 149–158. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2483604&tool=pmcentrez&rendertype=abstract

Kitson, A. L., Rycroft-Malone, J., Harvey, Gill, McCormack, Brendan, Seers, K., & Titchen, A. (2008). Evaluating the successful implementation of evidence into practice using the PARiHS framework: theoretical and practical challenges. *Implementation Science*, *3*. doi: 10.1186/1748-5908-3-1.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124–144. doi: 10.1037/a0017736

Landsverk, J., Brown, C. H., Rolls Reutz, J., Palinkas, L., & Horwitz, S. M. (2010). Design elements in implementation research: A structured review of child welfare and child mental health studies. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 54–63. doi: 10.1007/s10488-010-0315-y

Larsen, T., & Samdal, O. (2007). Implementing second step: Balancing fidelity and program adaptation. *Journal of Educational & Psychological Consultation*, *17*, 1–29. doi:10.1207/s1532768Xjepc1701_1

Latham, M., Myles, P., & Ricketts, T. (2003). *Development of a multiple choice questionnaire to measure changes in knowledge during CBT training*. Open paper. British Association of Behavioural and Cognitive Psychotherapy Annual Conference, York University.

Lavori, P. W., Rush, A. J., Wisniewski, S. R., Alpert, J., Fava, M., Kupfer, D. J.,… Trivedi, M. (2001). Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biological Psychiatry*, *50*, 792–801. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11720698

Lehman, W., Greener, J., & Simpson, D. (2002). Assessing organizational readiness for change. *Journal of Substance Abuse Treatment*, *22*, 197–209. doi:10.1016/S0740-5472(02)00233-7

Lewin, J., Lall, V., & Kratochwill, T. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology*, *49*, 55–79. doi:10.1016/j.jsp.2010.09.002

Lewis, C., Comtois, K., & Krimer, Y. (2011, March). *A comprehensive review of dissemination and implementation science instruments.* Poster presented at the annual meeting of the National Institutes of Health Dissemination and Implementation Conference, Bethesda, MD.

Lomas, J. (1993). Diffusion, dissemination and implementation: Who should do what? *Annals of the New York Academy of Sciences*, *703*, 226–237. doi:10.1111/j.1749-6632.1993.tb26351.x

Lopez, M., Osterberg, L., Jensen-Doss, A., & Rae, W. (2011). Effects of workshop training for providers under mandated use of evidence-based treatment. *Administration and Policy in Mental Health and Mental Health Services Research*. *38*(4), 301–312. doi: 10.1007/s10488-010-0326-8

Lovejoy, T. I., Demireva, P. D., Grayson, J. L., & McNamara, J. R. (2009). Advancing the practice of online psychotherapy: An application of Rogers' diffusion of innovations theory. *Psychotherapy: Theory, Research, Practice, Training*, *46*, 112–124. doi: 10.1037/a0015153

Luce, B., Kramer, J., Gooman, S., Connor, J., Tunis, S., Whicher, D., & Schwartz, J. (2009). Rethinking randomized clinical trials for comparative effectiveness research: The need for transformational change. *Annals of Internal Medicine*, *151*, 206–209.

Macpherson, H. (2004). Pragmatic clinical trials. *Complementary Therapies in Medicine*, *12*, 136–40. doi: 10.1016/j.ctim.2004.07.043

March, J. (May 2011). *Twenty years of comparative treatment trials in pediatric psychiatry*. University of Illinois at Chicago Grand Rounds, Chicago, Illinois.

March, J., Silva, S., Compton, S., Shapiro, M., Califf, R., & Krishnan, R. (2005). The case for practical clinical trials in psychiatry. *American Journal of Psychiatry*, *162*, 836–846. doi:10.1176/appi.ajp.162.5.836

Mays, N., & Pope, C. (2000). Assessing quality in qualitative research. *BMJ*, *320*, 50–52. Retrieved from http://www.bmj.com/content/320/7226/50.1.extract

McGraw, S. A., Sellers, D. E., Stone, E. J., & Bebchuk, J. (1996). Using process data to explain outcomes: An illustration from the child and adolescent trial for cardiovascular health (CATCH). *Evaluation Review*, *20*, 291–312. doi:10.1177/0193841X9602000304

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist*, *65*, 73–84. doi: 10.1037/a0018121

McLeod, B. (2001). *The therapy process observational coding system for child psychotherapy.* Unpublished manuscript, University of California, Los Angeles.

McLeod, B. D., & Weisz, J. R. (2010). The therapy process observational coding system for child psychotherapy-strategies scale. *Journal of Clinical Child and Adolescent Psychology 39*, 436–443. doi: 10.1080/15374411003691750

Meyer, G. (2004). Diffusion methodology: Time to innovate? *Journal of Health Communication*, *9*, 59–69. doi: 10.1080/10810730490271539

Miller, W., Yahne, C., Moyers, T., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*, *72*, 1050–1062. doi:10.1037/0022-006X.72.6.1050

Moore, G., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, *2*, 192–222. doi:10.1287/isre.2.3.192

Morgenstern, J., & McKay, J. R. (2007). Rethinking the paradigms that inform behavioral treatment research for substance use disorders. *Addiction*, *102*, 1377–1389. doi:10.1111/j.1360-0443.2007.01882.x

Moyers, T., Martin, T., Catley, D., Harris, K., & Ahluwalia, J. (2003). Assessing the integrity of motivational interviewing interventions: Reliability of the Motivational Interviewing Skills Code. *Behavioral and Cognitive Psychotherapy*, *31*, 177–184. doi:10.1017/S1352465803002054

Myles, P. J., Latham, M., & Ricketts, T. (2003). *The contributions of an expert panel in the development of a new measure of knowledge for the evaluation of training in cognitive behavioural therapy*. Open paper. British Association of Behavioural and Cognitive Psychotherapies Annual Conference, York University.

O'Connor, C., Small, S. A., & Cooney, S. M. (2007). *Program fidelity and adaptation: Meeting local needs without compromising program effectiveness.* What Works, Wisconsin Research to Practice Series, 4. Madison, WI: University of Wisconsin–Madison/Extension.

Ogles, B., Lunnen, K., Gillespie, D., & Trout, C. (1996). Conceptualization and initial development of the Ohio Scales. In C. Liberton,, K. Kutash,, & R. Friendman (Eds.), *The 8th Annual Research Conference Proceedings, A system of care for children's mental health: Expanding the research base* (pp. 33–37). Tampa, FL: University of South Florida, Florida Mental Health Institute, Research and Training Center for Children's Mental Health.

Okamura, K., Nakamura, B., Higa McMillan, C., Mueller, C., & Hayashi, K. (2010, November). *Psychometric evaluation of the Knowledge of Evidence-Based Questionnaire in a community sample of mental health therapists*. Poster presented at the annual meeting of the Association for Behavioral and Cognitive Therapies, San Francisco, CA.

Ozer, E. J., Wanis, M. G., & Bazell, N. (2010). Diffusion of school-based prevention programs in two urban districts: Adaptations, rationales, and suggestions for change. *Prevention Science*, *11*, 42–55. doi:10.1007/s11121-009-0148-7

Palinkas, L., Aarons, G., Horwitz, S., Chamberlain, P., Hurlburt, M., & Landsverk, J. (2011). Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 44–53. doi: 10.1007/s10488-010-0314-z

Palinkas, L., Schoenwald, S. K., Hoagwood, K., Landsverk, J., Chorpita, B. F., & Weisz, J. R. (2008). An ethnographic study of implementation of evidence-based treatments in child mental health: First steps. *Psychiatric Services*, *59*, 738–746. doi: 10.1176/appi.ps.59.7.738

Perpepletchikova, F., & Kazdin, A. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383. doi:10.1093/clipsy/bpi045

Pluye, P., Potvin, L., & Denis, J-L. (2004). Making public health programs last: Conceptualizing sustainability. *Evaluation and Program Planning*, *27*, 121–133. doi:10.1016/j.evalprogplan.2004.01.001

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A.,… Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. Administration and Policy in Mental Health and Mental Health Services Research, *38*(2), 65–76. doi: 10.1007/s10488-010-0319-7

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research*, *36*(1), 24–34. doi: 10.1007/s10488-008-0197-4

Rakovshik, S. G., & McManus, F. (2010). Establishing evidence-based training in cognitive behavioral therapy: A review

of current empirical findings and theoretical guidance. *Clinical Psychology Review*, *30*, 496–516. doi: 10.1016/j.cpr.2010.03.004

Ringwalt, C., Ennett, S. T., Vincus, A., & Simons-Rudolph, A. (2004). Students' special needs and problems as reason for the adaptation of substance abuse prevention curricula in the nation's middle schools. *Prevention Science*, *5*, 197–206. doi:10.1023/B:PREV.0000037642.40783.95

Rogers, E. (1995). *Diffusion of innovations*. New York: The Free Press.

Rogers, E. (2004). A prospective and retrospective look at the diffusion model. *Journal of Health Communication*, *9*, 13–19. doi: 10.1080/10810730490271449

Rohrbach, L. A., Graham, J. W., & Hansen, W. B. (1993). Diffusion of a school-based substance abuse prevention program: predictors of program implementation. *Preventive Medicine*, *22*, 237–260. doi:10.1006/pmed.1993.1020

Rush, A. (2001). *Sequenced Treatment Alternatives to Relieve Depression (STAR*D).* In: Syllabus and Proceedings Summary, American Psychiatric Association 154th Annual Meeting, New Orleans, LA, May 5–10, p. 182.

Scheirer, M. A. (2005). Is sustainability possible? A review and commentary on empirical studies of program sustainability. *American Journal of Evaluation*, *26*, 320–347. doi:10.1177/1098214005278752

Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services*, *52*, 1190–1197. doi:10.1176/appi.ps.52.9.1190

Schoenwald, S. K., Hoagwood, K. E., Atkins, M. S., Evans, M. E., & Ringeisen, H. (2010). Workforce development and the organization of work: The science we need. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*, 71–80. doi:10.1007/s10488-010-0278-z

Shediac-Rizkallah, M. C., & Bone, L. R. (1998). Planning for the sustainability of community-based health programs: Conceptual frameworks and future directions for research, practice and policy. *Health Education Research*, *13*, 87–108. doi:10.1093/her/13.1.87

Sholomskas, D., Syracuse-Siewert, G., Rounsaville, B., Ball, S., Nuro, K., & Carroll, K. (2005). We don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology*, *73*, 106–115. doi:10.1037/0022-006X.73.1.106

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). *Educational Evaluation and Policy Analysis*, *32*, 351–371. doi: 10.3102/0162373710373389

Stetler, C. (2001). Updating the Stetler Model of research utilization to facilitate evidence-based practice. *Nursing Outlook*, *49*, 272–279. doi:10.1067/mno.2001.120517

Stewart, R. E., & Chambless, D. L. (2010). Interesting practitioners in training in empirically supported treatments: Research reviews versus case studies. *Journal of Clinical Psychology*, *66*, 73–95. doi: 10.1002/jclp

Stirman, S., DeRubeis, R., Crits-Christoph, P., & Brody, P. (2003). Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *Journal of Consulting and Clinical Psychology*, *71*, 963–972. doi: 10.1037/0022-006X.71.6.963

Stumpf, R. E., Higa-McMillan, C. K., & Chorpita, B. F. (2009). Implementation of evidence-based services for youth: assessing provider knowledge. *Behavior Modification*, *33*, 48–65. doi: 10.1177/0145445508322625.

Torrey, W., Finnerty, M., Evans, A., & Wyzik, P. (2003). Strategies for leading the implementation of evidence-based practices. *Psychiatric Clinics of North America*, *26*, 883–897. doi: 10.1016/S0193-953X(03)00067-4

Trent, L., Buchanan, E., & Young, J. (2010, November). *Development and initial psychometric evaluation of the Measure of Disseminability (MOD).* Poster presented at the annual meeting of the Association for Behavioral and Cognitive Therapies, San Francisco, CA.

Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association*, *290*, 1624–1632. doi: 10.1001/jama.290.12.1624

Valente, T. W., & Davis, R. L. (1999). Accelerating the diffusion of innovations using opinion leaders. *Annals of the American Academy of Political and Social Science*, *566*, 55–67. doi: 10.1177/0002716299566001005

Webb, C., DeRubeis, R., & Barber, J. (2010). Clinician adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*, 200–211. doi:10.1037/a0018912

Weisz, J. R., Donenberg, G. R., Han, S. S., & Kauneckis, D. (1995). Child and adolescent psychotherapy outcomes in experiments versus clinics: Why the disparity? *Journal of Abnormal Child Psychology*, *23*, 83–106. doi:10.1007/BF01447046

West, S., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D., Holtgrave, D.,… Mullen, D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*, 1359–1366. doi:10.2105/AJPH.2007.124446

Young, J., & Beck, A. (1980). *The development of the Cognitive Therapy Scale.* Unpublished manuscript, University of Pennsylvania, Philadelphia, PA.

CHAPTER

6 Virtual Environments in Clinical Psychology Research

Nina Wong *and* Deborah C. Beidel

Abstract

Virtual environments (VEs) represent a new and powerful medium through which an individual can become immersed in an "alternative reality." Applications of these environments are increasingly used to treat psychological disorders such as anxiety and autism spectrum disorders. We review the available literature regarding the use of VEs to treat these and other clinical disorders, highlighting both what we know and what we need to learn. We end with suggestions for integrating VE into future research endeavors.

Key Words: Virtual environments, virtual reality, virtual exposure, treatment, clinical research

Since the introduction of computer-generated graphics in the early 1960s, the development and utilization of *virtual environments* (VEs) has flourished. Virtual reality technology was initially developed through investments by the federal government, military research, and NASA (Wiederhold & Wiederhold, 2005) and has been increasingly used for architecture and design, visualization of scientific models, education and training, entertainment, and medicine over the past two decades (Blade & Padgett, 2002).

Virtual reality (VR) is defined by the integration of computer graphics, body tracking devices, visual displays, and sensory input devices in real time to synthesize a three-dimensional computer-generated environment. The most commonly used approach to facilitate VR involves a head-mounted display (HMD) consisting of a visor with separate visual display screens for each eye. Perception of the actual surrounding environment is blocked by focusing on the visor's VE display screens. A body tracking device is connected to the HMD and matches the patient's VE to real-life head movements (i.e., if the patient turns to the right, the right side of the environment is displayed). In addition to the visual images, auditory, tactile, and olfactory stimuli are often included to increase patients' immersion in the VE. In essence, a virtual environment is simulated and can be controlled or deliberately manipulated such that individuals immerse themselves into lifelike experiences with surprising authenticity.

Although psychotherapy and clinical psychology research have only recently taken advantage of such technology (see Glantz, Rizzo, & Graap, 2003; Wiederhold & Wiederhold, 2005), the use of VR and computer-based therapies ranked 3rd and 5th, respectively, of 38 interventions predicted to increase by 2012 (Norcross, Hedges, & Prochaska, 2002). The growing enthusiasm for VE applications in clinical psychology merits a review of the current literature and suggestions for future research. This chapter begins by highlighting findings of VE technology in clinical psychology, organized by diagnostically relevant categories, and concludes by discussing the limitations of VEs in clinical psychology research and directions for integrating VEs into future research paradigms.

Virtual Environments for Anxiety Disorders

Efficacious behavioral treatment for anxiety disorders involves a systematic exposure to situations and stimuli that evoke fear. With repeated and prolonged exposure, the patient's anxiety responses gradually diminish through a process of habituation (Wolpe, 1958). Exposure-based treatments for anxiety disorders are well established and certainly the gold standard (Barlow, 2002; Chambless & Ollendick, 2001; Deacon & Abramowitz, 2004). Typical exposure modalities include imaginal or *in vivo* presentation of the feared stimulus. When patients have difficulty with imaginal exposure or the feared situations cannot be recreated *in vivo*, VEs may serve as a clinical tool to enhance stimulus presentation.

There has been a rapidly growing interest in the application of VR to treat anxiety disorders, with the appearance of a number of recent qualitative literature reviews (i.e., Anderson, Jacobs, & Rothbaum, 2004; Bush, 2008; Coelho, Waters, Hine, & Wallis, 2009; Gerardi, Cukor, Difede, Rizzo, & Rothbaum, 2010; Krijn, Emmelkamp, Ólafsson, & Biemond, 2004; Meyerbröker & Emmelkamp, 2010; Pull, 2005; Rothbaum & Hodges, 1999). Although the majority of studies focused on the treatment of specific phobias, such as aviophobia (fear of flying) and acrophobia (fear of heights), other studies examined the use of virtual reality exposure therapy (VRET) for social phobia or social anxiety, posttraumatic stress disorder, and panic disorder with or without agoraphobia. Collectively, VEs for the treatment of anxiety disorders have demonstrated promising results in case, comparative, and randomized controlled studies. Two meta-analytic reviews (Parsons & Rizzo, 2008; Powers & Emmelkamp, 2008) found large effect sizes for VRET, indicating that this intervention is highly effective in reducing anxiety and phobia symptoms (Parsons & Rizzo, 2008). Specifically, the average reduction in overall anxiety was 0.95 standard deviations, with the smallest effect size for PTSD (0.87) and largest for panic disorder with agoraphobia (1.79). In the second meta-analysis (Powers & Emmelkamp, 2008), VRET demonstrated a large overall mean effect size (Cohen's d = 1.11) relative to control conditions, and the effect was consistent across general measures of distress, as well as cognitive, behavioral, and psychophysiology measures (Powers & Emmelkamp, 2008). To inform future clinical research efforts incorporating VE technology in anxiety treatments, we next briefly review the utility of VR for the various anxiety disorders.

Specific Phobias

VEs have been used to successfully treat specific phobias, including aviophobia, acrophobia, and arachnophobia (fear of spiders). In particular, one of the most extensively utilized VEs was developed to treat fear of flying (Da Costa, Sardinha, & Nardi, 2008; Klein, 2000; Price, Anderson, & Rothbaum, 2008). Patients can be exposed to various flight experiences such as taxiing, takeoff, flight, and landing under calm and turbulent weather conditions through a virtual airplane environment. Speakers emitting low-frequency sound waves are built into a platform on which the patient sits and feels the vibrations associated with takeoff. With respect to efficacy, a randomized clinical trial (Rothbaum, Hodges, Smith, Lee, & Price, 2000) found that individuals in the VR exposure and standard *in vivo* exposure conditions reported a significant decrease in fear from pretreatment to posttreatment, whereas the waitlist control group did not show a significant change. Participants who received VRET or standard *in vivo* exposure maintained treatment gains at the 1-year follow-up (Rothbaum, Hodges, Anderson, Price, & Smith, 2002) and did not have a significant increase in fear of flying even after the September 11th attacks (Anderson et al., 2006). A second randomized control trial (Rothbaum et al., 2006) replicated and extended the study by adding a posttreatment flight as a behavioral outcome measure. In addition to decreases in self-reported measures of anxiety from pretreatment to posttreatment for both the standard *in vivo* and VRET groups, 76 percent of both treatment groups completed the posttreatment flight relative to 20 percent of the waitlist group, and both groups maintained treatment gains at the 6- and 12-month follow-ups (Rothbaum et al., 2006). Only one study directly compared VRET to computer-assisted psychotherapy for fear of flying (Tortella-Feliu et al., 2011). All three treatment conditions (VRET, computer-based treatment with therapist, and computer-based treatment without therapist) showed large within-group effect sizes and were equally effective at significantly reducing fear of flying at posttreatment and 1-year follow-up measures. Collectively, there appears to be substantial data to support the use of VE for fear of flying.

VE treatment shows preliminary promise for the treatment of other specific phobias. Two small studies used virtual environments to treat acrophobia. Glass elevators, footbridges, and outdoor balconies at varying heights are simulated in the VE. On average, participants treated with VR for height

phobia experienced decreases in symptoms at posttreatment relative to the waitlist group (Rothbaum et al., 1995). Another study (Coelho, Silva, Santos, Tichon, & Wallis, 2008) compared the effects of three VRET or *in vivo* exposure sessions for height phobia. Five patients received *in vivo* exposure to an eight-story hotel, and ten patients received exposure therapy to a virtual hotel. Both groups showed decreased anxiety and avoidance symptoms; however, patients appeared to habituate more quickly in the VE. In a treatment study for spider phobia, 83 percent of patients who received four 1-hour VR treatment sessions showed statistically and clinically significant improvement at posttreatment, relative to no changes in a waitlist control group (Garcia-Palacios, Hoffman, Carlin, Furness, & Botella, 2002). With regard to other specific phobias, a case study on the efficacy of VRET to treat driving phobia suggest that patients may no longer meet diagnostic criteria after treatment, and recommended using VRET to lower anxiety to the point of beginning *in vivo* exposure therapy (Wald & Taylor, 2003). In one small study, six patients with claustrophobia were treated with a multicomponent therapy program that included four sessions of VR (Malbos, Mestre, Note, & Gellato, 2008). At follow-up, their treatment gains generalized to other settings (i.e., using the elevator alone). Based on these collective findings, VRET appears to be an effective tool for the treatment of specific phobias. Interestingly, 76 percent of people with specific phobia or social phobia reported a preference for VRET over *in vivo* exposure (Garcia-Palacios, Botella, Hoffman, & Fabregat, 2007).

Social Phobia/Fear of Public Speaking

To date, VR has been successfully used to treat specific phobias in particular situations with powerful physical cues (e.g., distance cues for heights or strong vibrations and loud noises for flying). However, much more so than for specific phobias, therapists face significant challenges finding appropriate *in vivo* exposure contexts for individuals with social phobia (SP), which is characterized by a pattern of excessive fear of social situations or performances in which an individual may be scrutinized by others (American Psychiatric Association, 2000). The third most common mental health disorder in the United States, SP affects 8 percent of all youth and its prevalence ranges from 5 to 8 percent in the general population (Beidel & Turner, 2007).

Common distressful situations for people with SP include public speaking, meeting new people, and speaking during meetings or class (Ruscio et al., 2008; Turner, Beidel, & Townsley, 1992). However, a significant barrier to the treatment of SP lies in the difficulty of recruiting audience members to create *in vivo* social or public-speaking situations (Olfson et al., 2000). Thus, VEs have been recently developed as a possible alternative context for exposure therapy for social phobia (Klinger et al., 2003; Roy et al., 2003). VEs for social phobia, such as a virtual auditorium or conference room, primarily target public-speaking fears. Given that the hallmark of social phobia is a fear of negative evaluation by other people, the ability of the VE to elicit that fear is necessary for VR to work. Thus, the environment and the people in the environment have to feel realistic—cartoonish-looking avatars might not provide an environment reminiscent of the individual's actual fear. Fortunately, heightened physiological responses appear to be elicited in healthy controls in a VR speech task (Kotlyar et al., 2008). However, that same study observed similar increases in diastolic blood pressure, systolic blood pressure, and heart rate in participants completing an *in vivo* math task. These physiological findings should be interpreted as preliminary given that the VR speech task and *in vivo* math task may not be directly comparable, even though both may elicit substantial subjective and physiological distress.

A number of small studies investigated the efficacy of VRET for the nongeneralized subtype of social phobia, namely public-speaking fears. North and colleagues (1998) compared VRET to a control condition for participants with public-speaking phobia. The VRET condition was a large virtual audience, and the control condition was an unrelated neutral VE. The six participants who completed the VRET reported improvement on their attitude toward public speaking and subjective units of distress, while those in the control condition showed no change (North, North, & Coble, 1998). Improvement was also reported in a single-case study where VR public-speaking exposure therapy was part of a larger cognitive-behavioral therapy (CBT) program for two woman diagnosed with social phobia (Anderson, Rothbaum, & Hodge, 2003). Findings indicate that at posttreatment, patients' subjective rating of anxiety during the task and anxiety symptoms as measured by the Personal Report of Confidence as a Public Speaker Questionnaire (PRCS) decreased. Similar results were reported for an 8-week manualized treatment that included VRE with CBT (Anderson, Zimand, Hodges, & Rothbaum, 2005). For the ten

participants diagnosed with social phobia or panic disorder with a primary fear of public speaking, self-reported measures of anxiety and public-speaking fears decreased after treatment (Anderson et al., 2005). In addition, participants reported feeling satisfied by the treatment, and treatment gains were maintained at 3-month follow-up. However, since there were no control groups and VR was not tested in isolation in the two studies by Anderson and colleagues (2005), the efficacy of VR exposure alone cannot be confirmed at this time.

Several studies further examined public-speaking anxiety in students without a clinical diagnosis of SP (Harris, Kemmerling, & North, 2002; Lister, Piercey, & Joordens, 2010; Wallach, Safir, & Bar-Zvi, 2009). Relative to no treatment, VRET was more effective at reducing public-speaking fears among a small sample of undergraduate students who reported high levels of public-speaking fears (Harris et al., 2002). In that study, Harris and colleagues (2002) surveyed an introductory public-speaking class at a large university and recruited participants based on a PRCS cutoff score of more than 16. Eight participants in the VRET condition received exposure to a virtual auditorium, and six participants were in the waitlist control condition. At posttest, self-reported levels of confidence as a speaker were significantly different between the VRET condition and waitlist controls. Also using university students with public-speaking anxiety, a larger randomized clinical trial examined VR CBT as an alternative to CBT for public-speaking anxiety in an Israeli sample (Wallach et al., 2009). Eighty-eight participants with public-speaking anxiety were randomly assigned to VR CBT, CBT, and waitlist control. Relative to waitlist controls, both the VR CBT and CBT groups reported significantly lower anxiety on social anxiety and public-speaking questionnaires and lower anxiety during a 10-minute behavioral speech performance task (Wallach et al., 2009). Although that investigation was the only study to include a behavioral performance task, no significant differences were found on observer ratings of anxiety between the two treatment groups. While VR CBT was not superior to CBT in that study, twice as many participants dropped out from CBT than from VR CBT (Wallach et al., 2009).

Additional studies examined whether the changes in the virtual environment influenced public-speaking anxiety in students without a clinical diagnosis of SP (Pertaub, Slater, & Barker, 2002; Slater, Pertaub, Barker, & Clark, 2006). Pertaub and colleagues (2002) recruited university students and staff to give a speech to either a neutral, positive, or negative virtual audience. In that study, all three conditions elicited anxiety in participants with elevated PRCS scores. All participants reported feeling anxious when giving a speech to the negative audience regardless of their PRCS scores (Pertaub et al., 2002). Slater and colleagues (2006) later replicated the study by Pertaub and colleagues (2002) by comparing participants' physiological responses and subjective levels of distress when speaking to an empty VR auditorium or a VR audience. Participants were classified by PRCS scores as either confident speakers (PRCS $\geq$ 20) or phobic speakers (PRCS $\leq$ 10). Unlike confident speakers, phobic speakers reported higher levels of anxiety and somatic responses in both the empty VR auditorium and VR audience conditions. Physiologically, the phobic speakers showed a decreasing trend in heart rate when speaking to an empty VR room compared to those speaking to an audience, and increased arousal when the VR audience was present (Slater et al., 2006). These two studies suggest that participants respond to contextual changes in the VE.

Collectively, the above studies indicate that VEs elicit a response for participants with public-speaking fears. Even a three-dimensional video of a virtual audience presented on a standard CRT television elicits a performance fear response (Lister et al., 2010). In that study, Lister and colleagues (2010) similarly recruited undergraduate students from a general introduction psychology course, based on a PRCS cut-off score of more than 21. Nine participants in the VR condition stood behind a podium in front of the TV and were exposed to the virtual audience through the use of polarized shutter glasses. The speakers initially stood in front of the virtual audience when the audience paid no attention (2 minutes) and subsequently read a text while the audience paid attention (additional 2 minutes). Results showed that skin conductance and heart-rate measures increased after performing the public-speaking task to the VR audience, and subjective ratings of anxiety and negative self-beliefs about public-speaking ability decreased (Lister et al., 2010).

Despite the extant literature on VEs for public-speaking anxiety, VRET for the generalized subtype of social phobia remains much less studied. Currently, only one controlled study in France (Klinger et al., 2005) used VEs to target symptoms other than public-speaking fears. In that study, 36 participants clinically diagnosed with social phobia were assigned to either 12 sessions of VRET

or group CBT. Klinger and colleagues (2005) used exposure VEs that replicated a variety of social situations, including public speaking, short interpersonal interaction (e.g., small talk at a dinner conversation), assertiveness (e.g., having a viewpoint challenged), and evaluation (e.g., completing a task while being observed). After treatment, participants in both conditions showed increased global measures of functioning and reported decreased symptoms of social phobia, and the efficacy of VRET to the control traditional group CBT was not statistically different based on effect size comparisons (Klinger et al., 2005). However, these findings should be interpreted cautiously because that study did not include a third condition such as placebo or waitlist control. Future research on VRET for the generalized subtype of social phobia is needed.

Posttraumatic Stress Disorder

U.S. military deployment to Operation Iraqi Freedom/Operation Enduring Freedom has been extensive, and up to 18.5 percent of returning veterans are diagnosed with posttraumatic stress disorder (PTSD) (Hoge, Auchterlonie, & Milliken, 2006; Tanielian & Jaycox, 2008; Smith et al., 2008). Although veterans with PTSD are often reluctant to seek mental health services (Hoge et al., 2004) they tend to be higher users of medical care services (Kessler, 2000; Solomon & Davidson, 1997), and a majority (>90 percent) also seek disability compensation for debilitating occupational impairment. Unlike other anxiety disorders characterized by anticipatory fear, individuals with PTSD experience anxiety related to previous traumatic events that actually happened. Positive symptoms of PTSD include intrusive thoughts, re-experiencing, hyperarousal, and avoidance. Behavioral treatment for PTSD often relies on imaginal exposure because the traumatic events may be difficult or unethical to recreate. Recently, VEs have been developed to augment imaginal exposure for combat-rated PTSD with sensory cues such as visual, auditory, olfactory, and haptic stimuli (Cukor, Spitalnick, Difede, Rizzo, & Rothbaum, 2009). For example, at least one VR technology can recreate 13 scents from Middle Eastern wartime settings, including burned rubber, gunpowder, and body odor (Rizzo et al., 2010).

A few case studies examined the efficacy of VRET for combat-related PTSD (Reger & Gahm, 2008; Wood et al., 2007). Reger and Gahm (2008) treated one active-duty U.S. military soldier diagnosed with combat-related PTSD. The treatment consisted of six, 90-minute VRET sessions over the course of 4 weeks as part of a larger treatment protocol including psychoeducation, relaxation training, and *in vivo* exposure. Although his score on a self-report PTSD inventory decreased at posttreatment, and gains were maintained at 7-week follow-up, the specific effects of VRET cannot be isolated from the multicomponent treatment. Similar findings were reported by Wood and colleagues (2007)—one veteran reported decreases in symptoms of combat-related PTSD and physiological arousal following VRET. Although the ability to generalize findings from these two case studies is limited, a pilot study found that 12 men with PTSD reported decreased levels of PTSD and depression after 20 sessions of VRET, and 75 percent of participants no longer met criteria for PTSD (Wood et al., 2009). Furthermore, a larger study (Reger et al., 2011) examined the effectiveness of VRET for 24 active-duty soldiers who sought treatment following a deployment to Iraq or Afghanistan. Patients reported a significant reduction of PTSD symptoms as measure by the PTSD Checklist (Military version) at posttreatment.

VR has also been used to treat civilian-related PTSD. Two small studies by one research group (Difede, Cukor, Patt, Giosan, & Hoffman, 2006; Difede et al., 2007) investigated the utility of VRET for civilian and disaster workers who were directly exposed to the World Trade Center attacks on September 11, 2011 and diagnosed with PTSD. Both studies reported that participants who received VR treatment showed a significant decrease in PTSD symptoms relative to the waitlist control group, and improvements were maintained at 6-month follow-up (Difede et al., 2007).

Panic Disorder With or Without Agoraphobia

Panic disorder with or without agoraphobia is associated with substantial severity and impairment, and lifetime prevalence estimates are 1.1 percent and 3.7 percent, respectively (Kessler et al., 2006). Individuals with panic disorder report significant distress over panic attacks. Panic attacks are characterized by the sudden and unexpected onset of a period of intense fear and discomfort with a cluster of physical and cognitive symptoms (American Psychiatric Association, 2000). Patients with panic disorder are often concerned about the implications of their panic attacks and report a persistent concern and avoidance of future attacks. Panic disorder may occur with or without agoraphobia. Agoraphobia

is characterized by severe anxiety and avoidance of situations in which a panic attack might occur and fear that it might be difficult or embarrassing to escape (e.g., crowds, public transportation, traveling alone, etc.). Individuals with panic disorder, with or without agoraphobia, often report intense fear and need to escape, in addition to a number of physical sensations in their body such as heart palpitations, difficulty breathing, feeling unsteady or nauseated, and trembling.

Although interoceptive therapy remains undisputed as one of the gold-standard treatments for panic disorder, currently at least one VE exists to augment interoceptive exposure. The VE technology incorporates external audio/visual stimuli with interoceptive cues to trigger bodily sensations, such as blurred vision and audible rapid heartbeats, while the person is in a virtual bus or tunnel (Botella et al., 2004, 2007). Recently, the utility of VRET for panic disorder with or without agoraphobia has been examined in a few small controlled studies (Botella et al., 2007; Pérez-Ara et al., 2010). In one study, patients with panic disorder, with or without agoraphobia, who received either *in vivo* exposure or VRET as part of a multicomponent treatment, improved similarly and showed more improvement than waitlist controls at posttreatment (Botella et al., 2007). Both treatment groups reported decreased catastrophic thoughts and improvements on clinical global impression scores with treatment gains maintained at 12-month follow-up. Interestingly, that study reported that 100 percent of the *in vivo* condition participants no longer had panic or had a 50 percent reduction in panic frequency at posttreatment, while the rates were 90 percent in the VRET condition and 28.57 percent in the waitlist control group (Botella et al., 2007). Finally, both treatment groups also reported similar levels of satisfaction. Another small study from the same research group compared treatment outcomes for patients with panic disorder with or without agoraphobia who received a multicomponent CBT program, including exposure to an agoraphobic VE (e.g., a crowded mall or narrow tunnel) in two conditions (Pérez-Ara et al., 2010). One condition received simultaneously presented audio and visual effects (e.g., audible rapid heartbeat or double vision). The other condition received the same VE for 25 minutes followed by traditional interoceptive exercises (e.g., head spinning and hyperventilation). Outcome did not differ by treatment condition; both treatments reported decreased fear, avoidance, and panic disorder severity at posttreatment, and gains were maintained at 3-month follow-up. In another study (Vincelli et al., 2003), VR was integrated into a multicomponent CBT strategy (Experiential Cognitive Therapy [ECT]) for the treatment of panic disorders with agoraphobia. The small controlled study compared treatment outcome in three groups, including an ECT condition in which participants received eight sessions of VR-assisted CBT, a CBT condition in which they received 12 sessions of traditional cognitive-behavioral approach, and a waitlist control condition. At posttreatment, both treatment groups reported significantly less frequent panic attacks and decreased levels of anxiety and depression (Vincelli et al., 2003). Similar findings in a larger study (Choi et al., 2005) reported that patients with panic disorder and agoraphobia improved at posttreatment, regardless if they received brief ECT (4 sessions) or 12 sessions of traditional panic treatment.

Virtual Environments for Developmental Disorders and Intellectual Disabilities

In addition to the anxiety disorders, VEs have been widely implemented for youth with developmental disorders, particularly autism spectrum disorders (ASDs). Given the neurodevelopmental challenges faced by children with these disorders, several potential concerns regarding the implementation of VE have been identified (Andersson, Josefsson, & Pareto, 2006; Parsons, 2005)—specifically, would adolescents with ASDs (a) be able to use the VEs appropriately, (b) understand the VEs as representational devices, and (c) learn new information from the VEs about social skills (Parsons, 2005)? Over the course of three investigations (Parsons, 2005), data indicated that individuals with ASDs can complete tasks, understand social situations, and learn social conventions through the VE, thus confirming the benefits of immersive VEs for children and adolescents with ASDs (Wallace et al., 2010).

For youth with ASDs, the use of VEs have primarily targeted social skill deficits (Cheng, Chiang, Ye, & Cheng, 2010; Mitchell, Parsons, & Leonard, 2007). Mitchell and colleagues (2007) created a virtual café wherein six adolescents with ASDs were able to practice their social skills anonymously (i.e., determining where to sit: either at an empty table or an empty seat with others already sitting, or were given the choice to ask others if a seat was available). After engaging in the VE experience, some adolescents showed greater levels of social understanding as measured by their ability to justify where they would sit and why they chose that seat. Interestingly,

participants were also able to generalize learned behavior to different VE contexts (i.e., from a virtual café to a live video of a bus). Another small study used a collaborative virtual learning environment to teach the use and comprehension of empathy (e.g., kindness, tolerance, and respect) to three children with ASDs (Cheng et al., 2010). Each participant showed improvement on the Empathy Rating Scale posttreatment and continued to maintain gains at follow-up relative to baseline. Results preliminarily suggest that the virtual learning environment system may be helpful in increasing empathic understanding among children with ASDs.

VEs have been used to target other skill deficits among youth with ASDs, such as crossing the street (Josman, Ben-Chaim, Friedrich, & Weiss, 2008). Additionally, a VR-tangible interaction, involving both virtual and physical environments, was used to address sensory integration and social skills treatment for 12 children with ASDs (Jung et al., 2006). Although not empirically studied, Wang and Reid (2009) also discussed the potential for VR technology in the cognitive rehabilitation of children with autism.

In addition to youth with ASDs, VEs have been used for people with intellectual disabilities (Standen & Brown, 2005; Stendal, Balandin, & Molka-Danielsen, 2011). VEs have the potential to teach independent living skills, including grocery shopping, food preparation, spatial orientation (Mengue-Topio, Courbois, Farran, & Sockeel, 2011), road safety, and vocational training (Standen & Brown, 2005). To enhance usability for a population with intellectual disabilities, Lannen and colleagues (2002) recommended the development of new devices based on existing data, and additional prototype testing of devices among individuals with learning disabilities.

VEs Used with Other Clinical Disorders

In addition to anxiety disorders and developmental disorders, VE applications are increasingly used in other clinical contexts. For example, VEs have been used to assess reactivity and cravings for alcohol (Bordnick et al., 2008) and other substances (Bordnick et al., 2009; Culbertson et al., 2010; Girard, Turcotte, Bouchard, & Girard 2009; Traylor, Bordnick, & Carter, 2008). In these scenarios, participants walk through a virtual party where alcohol or other substance-related cues are available. As they encounter different cues (a bottle of beer, a bartender, other people drinking and smoking), their cravings for that substance are assessed. In this way, changes in cue reactivity as a result of an intervention may be measured. VEs for eating disorders are designed with food-related and body-image cues to elicit emotional responses, and have been used to assess body-image distortion and dissatisfaction. The scenarios incorporate cues such as a virtual kitchen with high-calorie foods and scales showing the patients' real weight (Ferrer-García & Gutiérrez-Maldonado, 2005; Gutiérrez-Maldonado, Ferrer-García, Caqueo-Urízar, & Letosa-Porta, 2006; Gutiérrez-Maldonado, Ferrer-García, Caqueo-Urízar, & Moreno, 2010; Perpiñá, Botella, & Baños, 2003). However, it is not always necessary for VEs to present cue-related stimuli in order to assess psychological symptoms. For example, neutral VEs have been used to assess unfounded persecutory ideation among individuals on a continuum of paranoia (Freeman, Pugh, Vorontsova, Antley, & Slater, 2010). Unlike real life, where the behaviors of other people are never under complete experimental control, in VEs the behaviors of others are scripted by the designer. Thus, VEs can provide a high level of experimental control. The VE avatars do only what they are programmed to do, allowing a unique opportunity to observe how someone interprets their behavior. For example, an avatar that looks in the direction of the person immersed in the environment may be interpreted as a "curious glance" by someone with no evidence of psychosis or as a "menacing stare" by an individual with paranoia. Finally, cognitive-based assessment and training may be embedded into VEs. For example, cognitive training programs in a VE have been used for older adults with chronic schizophrenia (Chan, Ngai, Leung, & Wong, 2010), and continuous performance tests in a VE have been used for youth with attention-deficit/hyperactivity disorders (Pollak et al., 2009; Rizzo et al., 2000). Collectively, these studies suggest that VEs may be used broadly across a wide range of clinical disorders.

Limitations of VEs in Clinical Psychology Research

There are several challenges for the integration of VEs into research and clinical settings (Blade & Padgett, 2002). We will discuss issues that may affect the utility and efficacy of VE. First, for VEs to be effective, several basic conditions are necessary (Foa & Kozak, 1986), including active participation and immersion in the environment (Slater, Pertaub, & Steed, 1999), generalizability to real-life situations, and the ability to elicit physiological responses (North, North, & Coble, 1997/1998;

Schuemie, van der Straaten, Krijn, & van der Mast, 2000). Among these conditions, the user's level of immersion or presence is usually described by the quality of the VE experience, and may be directly related to the efficacy of the experience (Wiederhold & Weiderhold, 2005). Overall, VEs do seem to provide the level of immersion needed for treatment efficacy (Alsina-Jurnet, Gutiérrez-Maldonado, & Rangel-Gómez, 2011; Gamito et al., 2010; IJsselsteijn, de Ridder, Freeman, & Avons, 2000; Krijn, Emmelkamp, Biemond, et al., 2004; Krijn, Emmelkamp, Ólafsson, Schuemie, & van der Mast, 2007; Price & Anderson, 2007; Price, Mehta, Tone, & Anderson, 2011; Villani, Riva, & Riva, 2007; Wallace et al., 2010; Witmer & Singer, 1998), thus making them a potentially valuable tool to augment exposure therapy and exposure research, particularly in instances where *in vivo* exposure is not feasible and the participant cannot produce a vivid imaginal scene.

A second issue that merits consideration is the potential for side effects due to being immersed in the VE environment, such as cyber sickness (Bruck & Watters, 2009). Although the increasing sophistication of the HMDs and tracking devices has dramatically decreased the likelihood of motion sickness in VEs, there is still a need to carefully evaluate patients both before and after the session to evaluate motion or simulator sickness symptoms such as nausea, dizziness, and headache. In our clinic, we advise patients to have a light meal about an hour prior to the treatment sessions to reduce the likelihood of side effects. Furthermore, we do not allow patients to leave the treatment clinic until any symptoms have dissipated. Although motion sickness is rare, it is necessary to evaluate.

Third, even though sophisticated environments exist, researchers and clinicians may be less likely to use this technology because of the time and effort required to learn proper equipment use. Although a degree in engineering or computer science is not necessary, some comfort with basic electronics such as dual-screen computer monitors, HMDs, audio amplifiers, and air compressors (in the case of olfactory stimulation) is beneficial. As with any type of equipment, technical difficulties are possible and require the ability to troubleshoot the problem (Segal, Bhatia, & Drapeau, 2011). The sophisticated VR units and specialized hardware required also remain costly (Gregg & Terrier, 2007) even though prices have dropped over the past decade (Rizzo, Reger, Gahm, Difede, & Rothbaum, 2009).

Despite these potential challenges, the benefits of VE in research and clinical contexts appear to outweigh the limitations (Glantz et al., 2003). As the efficacy data for VEs increase, additional research and demand for the intervention will increase, thereby making the purchase of VR compatible equipment a better investment for researchers, therapists, and patients alike. For example, VE appears to be a cost-effective treatment strategy (Wood et al., 2009)—the cost of VRET for 12 participants was estimated to be $114,490 less than the cost of treatment as usual ($439,000). Findings from an Internet survey on psychotherapists' perception regarding the use of VR treatments found that one of the highest-rated benefits was the ability to expose patients to stimuli that would otherwise be impractical or difficult to access (Segal et al., 2011). Another highly rated benefit was that therapists believe they have increased control over the situation. Indeed, VR has appealing features such as the potential to deliberately control what is presented to the client, the ability to tailor treatment environments to the needs of each individual within the confines of the technology, the ability to expose the client to a range of conditions that would be impractical or unsafe in the real world, and the ability to provide an alternative to individuals who are concerned about confidentiality or being seen in treatment by others (Glantz et al., 2003).

Integrating VE into Research Paradigms

As indicated above, VEs are now an accepted tool in the clinical treatment of individuals with anxiety disorders, and are emerging as useful for other disorders as well. We now turn our attention to the utility of VEs in research endeavors.

One of the clear advantages of using VEs in the research setting is the ability to assess behavior under standardized conditions. For example, consider the assessment of public-speaking anxiety. Previously, investigators who wanted to conduct a behavioral assessment faced the choice of having the individual deliver a speech to a camera or to a very small audience or to spend extensive time trying to arrange for a more realistic-size audience (usually more than five people) in order to conduct the assessment. When working with youth, there are additional issues associated with potentially using the youth's peers for a legitimate social-exposure task. With VE, the audience can be credibly simulated so there is no need to try and find a live audience. In addition to having a readily available virtual audience, the customizable nature of VEs allow the investigator to have full

control over the behavior of the "audience." Avatars can be activated to utter the same phrase or make the same movement at the same time for each assessment and for each research participant. Audience feedback can be specifically programmed to occur at the same time for every participant. Similarly, for substance abuse/dependence research, assessment of craving and other behaviors related to addiction can be measured using controlled environments. The ability to replicate the assessment before and after treatment, for example, allows for more calibrated and standardized behavioral assessments. In short, conducting behavioral assessments *in vivo* can be cumbersome and allows only limited controllability of the environment. Thus, the customizable and controllable features available through the application of VEs are potential solutions for researchers who want to fully control the environment in a standardized research protocol.

Additionally, as the protection of human subjects and patient confidentiality are top priorities for Institutional Review Board regulations, both researchers and participants can benefit from VE protocols in the laboratory, where patient confidentiality is easily maintained. Furthermore, although the use of audio/video recording and physiological assessments during the *in vivo* encounter is difficult to engineer, it is much more feasible through a VE. For example, when the distressful situation/event can be created in the clinic through the use of VE, there is the added advantage of being able to assess physiological responses (e.g., heart rate, electrodermal activity, etc.) and the ability to record the assessment, allowing for the coding of overt behaviors.

A disadvantage of VE in the assessment of psychopathology is that although the environments are customizable, they are unlikely to be able to exactly recreate every specific anxious situation that might exist. In the cases of most anxiety disorders, this is not an issue, as most individuals with anxiety disorders fear something that *might* happen. For example, individuals with a specific phobia of flying are afraid that the plane *might* crash. Thus, the VE must encompass visuals of the terminal, the runway, the inside of the plane—but because the event has not happened, the individual is able to accept the VE as is (color of the seats in the plane, number of people in the terminal). Similarly, for individuals with social phobia who are anxious in a job interview, the sex or race of the human resources director may not matter. In contrast, VE may actually limit the assessment of psychopathology among individuals with PTSD for whom the traumatic event has actually happened. In our experience, individuals with PTSD, when placed in a VE for assessment, will find any deviation from the actual traumatic event to be a distraction. Thus, if the actual event involved a collision with a dump truck, the individual with PTSD will resist immersion in a VE if the collision involves a tractor-trailer. They become difficult to engage in the VE because they are distracted by the "wrong elements." In such cases, VEs serve as a distinct disadvantage, not an advantage.

There are two groups of researchers that engage in the use of VE. The first group consists of individuals who either design/develop or partner with software engineers/companies in the design and development of software to be used in clinical and research settings. In many instances, VEs are developed by researchers in partnership with small businesses that specialize in VR, with funding provided by, for example, the Department of Defense or the National Institute of Health's (NIH) Small Business Innovation Research (SBIR)/Small Business Technology Transfer (STTR) grant mechanisms. Whereas small businesses are eligible to submit grants through the SBIR mechanism, the STTR mechanism requires a formal partnership between a small business and a research university. Other grant mechanisms, such as the R (Research) mechanisms, may be used for research when existing VEs are used for the purpose of other research questions. Individuals involved in the development come from many different backgrounds, including psychology, engineering, computer programming and gaming, and even art.

The second group of individuals who use VE are not primarily involved in its development, but the majority are psychologists who purchase/use software from companies designed specifically for that purpose. The largest hurdle in establishing a VE laboratory is the initial outlay for the equipment and software. The prices for systems typically used in research settings vary, but the cost for initial outlays ranges from $15,000 to $25,000 for hardware and software programs. Once past this initial investment, there are few additional costs (e.g., repair/replacement of broken equipment, maintenance contracts, software upgrades).

Although most of the interfaces are fairly intuitive, they are not turnkey. Thus, there is an initial learning curve for researchers to learn how to operate the system seamlessly in order to provide the most realistic experience. Individuals who are more familiar with videogames often find the system easier to use. Older adults, or those with less gaming experience, will require, and should be given,

practice so that they are familiar with the controls prior to participating in the research study.

The equipment required for VE is quite variable and depends on the type of media. Mixed-media environments can be sophisticated and may involve the use of "green screens" such as those used in the motion picture industry to allow the individual a more immersive experience. The setup requires a larger space so that the individual may walk around the green-screen environment, which enhances the experience. More commonly, VE systems use a desktop computer, dual monitors, an HMD, earphones, and perhaps a scent machine if olfactory cues are included. One distinct advantage is that the equipment is more compact, allowing it to be used in a typical "office"-size room. There is no opportunity for the individual to actually walk through the environment as he or she is tethered by wires to the computer that delivers the VE program. Thus, the individual is either seated or stands and moves through the environment via a game controller. Wearing the HMD can be somewhat cumbersome, and may not be feasible for children given its size and weight. However, as technology continues to improve, VE environments are changing. HMDs may be replaced by newer systems that require only the use of a video monitor.

As noted above, designing and developing VEs require collaboration among different disciplines. Clinical psychologists often provide the vision of what the software must be able to do in order to be useful for research, clinical, and data-collection purposes. Human factors psychologists are integral for ensuring that the hardware/software interface is user-friendly. Computer programmers and software engineers are responsible for writing the software. Artists/graphic designers design the visual layout/feel of the VE and build any avatars that populate the environment. Constructing VEs can be quite costly. Currently, our laboratory collaborates with a VE company to design VEs for children with social phobia. The design of one environment (various areas within an elementary school) and eight speaking avatars, plus the related background, will cost approximately $250,000 and take over 18 months to develop. This includes the conceptual storyboard design and conversational content of all avatar characters as well as the actual visual design and engineering.

Future Directions

Given the rapid technological progress of VEs for clinical psychology these past three decades, the extant literature remains limited by small sample sizes, reliance upon self-report data, measures that lack psychometric properties, and inconsistent methodological procedures. Future research will benefit from larger sample sizes, the use of clinical measures with better psychometric properties for the diagnostic assessment, behavioral measures with independent and blinded raters, randomized treatment and control groups, standardized VR treatment protocols across investigations (i.e., the number, duration, and frequency of sessions vary greatly), and the distinction between clinical and statistical significance (i.e., the number of patients who no longer meet diagnostic criteria versus patients who returned to normal levels of functioning) (see Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). In addition, relative to the literature on VEs for adults with anxiety disorders, the utility of VEs for anxious youth is virtually nonexistent (Bouchard, 2011).

For researchers, VEs offer flexibility and the ability to conduct standardized behavioral assessments that are realistic yet under the control of the investigator. The initial adoption of VEs as a research tool can be daunting and expensive, yet the advantages are many and the obstacles are relatively easy to overcome. As technology improves and individuals' familiarity/sophistication with virtual worlds becomes standard, research participants and clinical patients alike may be increasingly receptive to the use of VEs in both research and clinical settings.

Conclusion

VE systems have been most widely used for anxiety disorders, developmental disorders, and a number of other health-related disorders. To date, data are limited but VEs appear to be a viable alternative to other exposure modalities. Clearly, VEs can be a tool for clinical psychology treatment and research when appropriately used as an enhancement—not a replacement—for empirically supported behavioral treatments (Glantz et al., 2003; Gorini & Riva, 2008).

References

Alsina-Jurnet, I., Gutiérrez-Maldonado, J., & Rangel-Gómez, M. (2011). The role of presence in the level of anxiety experienced in clinical virtual environments. *Computers in Human Behavior*, *27*, 504–512.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: APA.

Anderson, P., Jacobs, C. H., Lindner, G. K., Edwards, S., Zimand, E., Hodges, L., et al. (2006). Cognitive behavior therapy for fear of flying: Sustainability of treatment gains after September 11. *Behavior Therapy*, *37*, 91–97.

Anderson, P., Jacobs, C., & Rothbaum, B. O. (2004). Computer-supported cognitive behavioral treatment of anxiety disorders. *Journal of Clinical Psychology*, *60*, 253–267.

Anderson, P., Rothbaum, B. O., & Hodges, L. F. (2003). Virtual reality exposure in the treatment of social anxiety: Two case reports. *Cognitive and Behavioral Practice*, *10*, 240–247.

Anderson, P. L., Zimand, E., Hodges, L. F., & Rothbaurn, B. O. (2005). Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure. *Depression and Anxiety*, *22*, 156–158.

Andersson, U., Josefsson, P., & Pareto, L. (2006). Challenges in designing virtual environments training social skills for children with autism. *International Journal on Disability and Human Development*, *5*, 105–111.

Barlow, D. H. (2002). *Anxiety and its disorders: The nature and treatment of anxiety and panic* (2nd ed.). New York: Guilford Press.

Beidel, D. C., & Turner, S. M. (2007). *Shy children, phobic adults: The nature and treatment of social anxiety disorder*. Washington, DC: American Psychological Association.

Blade, R. A., & Padgett, M. (2002). Virtual environments: History and profession. In K. M. Stanney & K. M. Stanney (Eds.), *Handbook of virtual environments: Design, implementation, and applications* (pp. 1167–1177). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bordnick, P. S., Copp, H. L., Traylor, A., Graap, K. M., Carter, B. L., Walton, A., & Ferrer, M. (2009). Reactivity to cannabis cues in virtual reality environments. *Journal of Psychoactive Drugs*, *41*, 105–112.

Bordnick, P. S., Traylor, A., Copp, H. L., Graap, K. M., Carter, B., Ferrer, M., & Walton, A. P. (2008). Assessing reactivity to virtual reality alcohol-based cues. *Addictive Behaviors*, *33*, 743–756.

Botella, C., Villa, H., García-Palacios, A., Baños, R., Perpiñá, C., & Alcañiz, M. (2004). Clinically significant virtual environments for the treatment of panic disorder and agoraphobia. *CyberPsychology and Behavior*, *7*, 527–535.

Botella, C. C., García-Palacios, A. A., Villa, H. H., Baños, R. M., Quero, S. S., Alcañiz, M. M., & Riva, G. G. (2007). Virtual reality exposure in the treatment of panic disorder and agoraphobia: A controlled study. *Clinical Psychology & Psychotherapy*, *14*, 164–175.

Bouchard, S. (2011). Could virtual reality be effective in treating children with phobias? *Expert Review of Neurotherapeutics*, 11, 207–213.

Bruck, S., & Watters, P. (2009). Cybersickness and anxiety during simulated motion: Implications for VRET. *Annual Review of CyberTherapy and Telemedicine*, *7*, 169–173.

Bush, J. (2008). Viability of virtual reality exposure therapy as a treatment alternative. *Computers in Human Behavior*, *24*, 1032–1040.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Chan, C. F., Ngai, E. Y., Leung, P. H., & Wong, S. (2010). Effect of the adapted virtual reality cognitive training program among Chinese older adults with chronic schizophrenia: A pilot study. *International Journal of Geriatric Psychiatry*, *25*, 643–649.

Cheng, Y., Chiang, H., Ye, J., & Cheng, L. (2010). Enhancing empathy instruction using a collaborative virtual learning environment for children with autistic spectrum conditions. *Computers & Education*, *55*, 1449–1458.

Choi, Y., Vincelli, F., Riva, G., Wiederhold, B. K., Lee, J., & Park, K. (2005). Effects of group experiential cognitive therapy for the treatment of panic disorder with agoraphobia. *CyberPsychology & Behavior*, *8*, 387–393.

Coelho, C. M., Silva, C. F., Santos, J. A., Tichon, J., & Wallis, G. (2008). Contrasting the effectiveness and efficiency of virtual reality and real environments in the treatment of acrophobia. *PsychNology Journal*, *6*, 203–216.

Coelho, C. M., Waters, A. M., Hine, T. J., & Wallis, G. (2009). The use of virtual reality in acrophobia research and treatment. *Journal of Anxiety Disorders*, *23*, 563–574.

Cukor, J., Spitalnick, J., Difede, J., Rizzo, A., & Rothbaum, B. O. (2009). Emerging treatments for PTSD. *Clinical Psychology Review*, *29*, 715–726.

Culbertson, C., Nicolas, S., Zaharovits, I., London, E. D., De La Garza, R., Brody, A. L., & Newton, T. F. (2010). Methamphetamine craving induced in an online virtual reality environment. *Pharmacology, Biochemistry and Behavior*, *96*, 454–460.

Da Costa, R. T., Sardinha, A., & Nardi, A. E. (2008). Virtual reality exposure in the treatment of fear of flying. *Aviation, Space, and Environmental Medicine*, *79*, 899–903.

Deacon, B. J., & Abramowitz, J. S. (2004). Cognitive and behavioral treatments for anxiety disorders: A review of meta-analytic findings. *Journal of Clinical Psychology*, *60*, 429–441.

Difede, J., Cukor, J., Jayasinghe, N., Patt, I., Jedel, S., Spielman, L., Giosan, C., & Hoffman, H. G. (2007). Virtual reality exposure therapy for the treatment of posttraumatic stress disorder following September 11, 2001. *Journal of Clinical Psychiatry*, *68*, 1639–1647.

Difede, J., Cukor, J., Patt, I., Giosan, C., & Hoffman, H. (2006). The application of virtual reality to the treatment of PTSD following the WTC attack. *Annals of the New York Academy of Sciences*, *1071*, 500–501.

Ferrer-García, M. M., & Gutiérrez-Maldonado, J. J. (2005). Assessment of emotional reactivity produced by exposure to virtual environments in patients with eating disorders. *Annual Review of CyberTherapy and Telemedicine*, *3*, 123–128.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin*, *99*, 20–35.

Freeman, D., Pugh, K., Vorontsova, N., Antley, A., & Slater, M. (2010). Testing the continuum of delusional beliefs: An experimental study using virtual reality. *Journal of Abnormal Psychology*, *119*, 83–92.

Gamito, P., Oliveira, J., Morais, D., Baptista, A., Santos, N., Soares, F., &… Rosa, P. (2010). Training presence: The importance of virtual reality experience on the "sense of being there." *Annual Review of CyberTherapy and Telemedicine*, *8,*103–106.

Garcia-Palacios, A., Hoffman, H., Carlin, A., Furness, T. A., & Botella, C. (2002). Virtual reality in the treatment of spider phobia: A controlled study. *Behaviour Research and Therapy*, *40*, 983–993.

Garcia-Palacios, A. A., Botella, C. C., Hoffman, H. H., & Fabregat, S. S. (2007). Comparing acceptance and refusal rates of virtual reality exposure vs. in vivo exposure by patients with specific phobias. *CyberPsychology & Behavior*, *10*, 722–724.

Gerardi, M., Cukor, J., Difede, J., Rizzo, A., & Rothbaum, B. (2010). Virtual reality exposure therapy for post-traumatic stress disorder and other anxiety disorders. *Current Psychiatry Reports*, *12*, 298–305.

Girard, B., Turcotte, V., Bouchard, S., & Girard, B. (2009). Crushing virtual cigarettes reduces tobacco addiction and treatment discontinuation. *CyberPsychology & Behavior*, *12*, 477–483.

Glantz, K., Rizzo, A., & Graap, K. (2003). Virtual reality for psychotherapy: Current reality and future possibilities. *Psychotherapy: Theory, Research, Practice, Training*, *40*, 55–67.

Gorini, A., & Riva, G. (2008). Virtual reality in anxiety disorders: The past and the future. *Expert Review of Neurotherapeutics*, *8*, 215–233.

Gregg, L., & Tarrier, N. (2007). Virtual reality in mental health: A review of the literature. *Social Psychiatry and Psychiatric Epidemiology*, *42*, 343–354.

Gutiérrez-Maldonado, J., Ferrer-García, M., Caqueo-Urízar, A., & Letosa-Porta, A. (2006). Assessment of emotional reactivity produced by exposure to virtual environments in patients with eating disorders. *CyberPsychology & Behavior*, *9*, 507–513.

Gutiérrez-Maldonado, J., Ferrer-García, M., Caqueo-Urízar, A., & Moreno, E. (2010). Body image in eating disorders: The influence of exposure to virtual-reality environments. *Cyberpsychology, Behavior, and Social Networking*, *13*, 521–531.

Harris, S. R., Kemmerling, R. L., & North, M. M. (2002). Brief virtual reality therapy for public speaking anxiety. *CyberPsychology & Behavior*, *5*, 543–550.

Hoge, C. W., Auchterlonie, J. L., & Milliken, C. S. (2006). Mental health problems, use of mental health services and attrition from military service after returning from deployment to Iraq or Afghanistan. *Journal of the American Medical Association*, *295*, 1023–1032.

Hoge, C. W., Castro, C. A., Messer, S. C., McGurk, D., Cotting, D. I., & Koffman, R. L. (2004). Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *New England Journal of Medicine*, *351*, 13–22.

IJsselsteijn, W. A., de Ridder, H., Freeman, J., & Avons, S. (2000). Presence: Concept, determinants and measurement. *Proceedings of the SPIE, Human Vision and Electronic Imaging V*, *3959*, 520–529.

Josman, N., Ben-Chaim, H., Friedrich, S., & Weiss, P. (2008). Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism. *International Journal on Disability and Human Development*, *7*, 49–56.

Jung, K., Lee, H., Lee, Y., Cheong, S., Choi, M., Suh, D., Suh, D., Oah, S., Lee, S., & Lee, J. (2006). The application of a sensory integration treatment based on virtual reality-tangible interaction for children with autistic spectrum disorder. *PsychNology Journal*, *4*, 145–159.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 285–299.

Kessler, R. C. (2000). Posttraumatic stress disorder: The burden to the individual and to society. *Journal of Clinical Psychiatry*, *61* (Suppl 5), 4–14.

Kessler, R. C., Chiu, W., Jin, R., Ruscio, A., Shear, K., & Walters, E. E. (2006). The epidemiology of panic attacks, panic disorder, and agoraphobia in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *63*, 415–424.

Klein, R. A. (2000). Virtual reality exposure therapy in the treatment of fear of flying. *Journal of Contemporary Psychotherapy, 30*, 195–207.

Klinger, E., Bouchard, S. S., Légeron, P. P., Roy, S. S., Lauer, F. F., Chemin, I. I., & Nugues, P. P. (2005). Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *CyberPsychology & Behavior*, *8*, 76–88.

Klinger, E., Chemin, I., Légeron, P., et al. (2003). Designing virtual worlds to treat social phobia. In B. Wiederhold, G. Riva, & M. D. Wiederhold (Eds.), *Cybertherapy 2003* (pp. 113–121). San Diego, CA: Interactive Media Institute.

Kotlyar, M., Donahue, C., Thuras, P., Kushner, M. G., O'Gorman, N., Smith, E. A., & Adson, D. E. (2008). Physiological response to a speech stressor presented in a virtual reality environment. *Psychophysiology*, *45*, 1034–1037.

Krijn, M., Emmelkamp, P. G., Biemond, R., de Wilde de Ligny, C., Schuemie, M. J., & van der Mast, C. G. (2004). Treatment of acrophobia in virtual reality: The role of immersion and presence. *Behaviour Research & Therapy*, *42*, 229–239.

Krijn, M. M., Emmelkamp, P. G., Ólafsson, R. P., & Biemond, R. R. (2004). Virtual reality exposure therapy of anxiety disorders: A review. *Clinical Psychology Review*, *24*, 259–281.

Krijn, M. M., Emmelkamp, P. G., Ólafsson, R. P., Schuemie, M. J., & van der Mast, C. G. (2007). Do self-statements enhance the effectiveness of virtual reality exposure therapy? A comparative evaluation in acrophobia. *CyberPsychology & Behavior*, *10*, 362–370.

Lannen, T. T., Brown, D. D., & Powell, H. H. (2002). Control of virtual environments for young people with learning difficulties. *Disability and Rehabilitation: An International, Multidisciplinary Journal*, *24*, 578–586.

Lister, H. A., Piercey, C., & Joordens, C. (2010). The effectiveness of 3-D video virtual reality for the treatment of fear of public speaking. *Journal of CyberTherapy and Rehabilitation*, *3*, 375–381.

Malbos, E. E., Mestre, D. R., Note, I. D., & Gellato, C. C. (2008). Virtual reality and claustrophobia: Multiple components therapy involving game editor virtual environments exposure. *CyberPsychology & Behavior*, *11*, 695–697.

Mengue-Topio, H., Courbois, Y., Farran, E. K., & Sockeel, P. (2011). Route learning and shortcut performance in adults with intellectual disability: A study with virtual environments. *Research in Developmental Disabilities*, *32*, 345–352.

Meyerbröker, K., & Emmelkamp, P. G. (2010). Virtual reality exposure therapy in anxiety disorders: A systematic review of process-and-outcome studies. *Depression and Anxiety*, *27*, 933–944.

Mitchell, P., Parsons, S., & Leonard, A. (2007). Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*, 589–600.

Norcross, J., Hedges, M., & Prochaska, J. (2002). The face of 2010: A Delphi poll on the future of psychotherapy. *Professional Psychology: Research and Practice*, *33*, 316–322.

North, M., North, S., & Coble, J. R. (1998). Virtual reality therapy: An effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, *3*, 2–6.

North, M. M., North, S. M., & Coble, J. R. (1997/1998). Virtual reality therapy: An effective treatment for psychological disorders. In G. Riva & G. Riva (Eds.), *Virtual reality in neuro-psycho-physiology: Cognitive, clinical and methodological issues in assessment and rehabilitation* (pp. 59–70). Amsterdam Netherlands: IOS Press.

Olfson, M., Guardino, M., Struening, E., Schneier, F. R., Hellman, F., & Klein, D. F. (2000). Barriers to the treatment of social anxiety. *American Journal of Psychiatry*, *157*, 521–527.

Parsons, S. (2005). Use, understanding and learning in virtual environments by adolescents with autistic spectrum disorders. *Annual Review of CyberTherapy and Telemedicine*, *3*, 207–215.

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*, 250–261.

Pérez-Ara, M. A., Quero, S. S., Botella, C. C., Baños, R. R., Andreu-Mateu, S. S., García-Palacios, A. A., & Bretón-López, J. J. (2010). Virtual reality interoceptive exposure for the treatment of panic disorder and agoraphobia. *Annual Review of CyberTherapy and Telemedicine*, *8*, 61–64.

Perpiñá, C. C., Botella, C. C., & Baños, R. M. (2003). Virtual reality in eating disorders. *European Eating Disorders Review*, *11*, 261–278.

Pertaub, D.P., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, *11*, 68–78.

Pollak, Y., Weiss, P. L., Rizzo, A. A., Weizer, M., Shriki, L., Shalev, R. S., & Gross-Tsur, V. (2009). The utility of a continuous performance test embedded in virtual reality in measuring ADHD-related deficits. *Journal of Developmental and Behavioral Pediatrics*, *30*, 2–6.

Powers, M. B., & Emmelkamp, P. G. (2008). Review: Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders*, *22*, 561–569.

Price, M., & Anderson, P. (2007). The role of presence in virtual reality exposure therapy. *Journal of Anxiety Disorders*, *21*, 742–751.

Price, M., Anderson, P., & Rothbaum, B. (2008). Virtual reality as treatment for fear of flying: A review of recent research. *International Journal of Behavioral Consultation & Therapy*, *4*, 340–347.

Price, M., Mehta, N., Tone, E. B., & Anderson, P. L. (2011). Does engagement with exposure yield better outcomes? Components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia. *Journal of Anxiety Disorders*, *25*, 763–770.

Pull, C. B. (2005). Current status of virtual reality exposure therapy in anxiety disorders. *Current Opinion in Psychiatry*, *18*, 7–14.

Reger, G. M., & Gahm, G. A. (2008). Virtual reality exposure therapy for active duty soldiers. *Journal of Clinical Psychology*, *64*, 940–946.

Reger, G. M., Holloway, K. M., Candy, C., Rothbaum, B. O., Difede, J., Rizzo, A. A., & Gahm, G. A. (2011). Effectiveness of virtual reality exposure therapy for active duty soldiers in a military mental health clinic. *Journal of Traumatic Stress*, *24*, 93–96.

Rizzo, A., Difede, J., Rothbaum, B. O., Reger, G., Spitalnick, J., Cukor, J., & Mclay, R. (2010). Development and early evaluation of the virtual Iraq/Afghanistan exposure therapy system for combat-related PTSD. *Annals of the New York Academy of Sciences*, *1208*, 114–125.

Rizzo, A., Reger, G., Gahm, G., Difede, J., & Rothbaum, B. O. (2009). Virtual reality exposure therapy for combat-related PTSD. In P. J. Shiromani, T. M. Keane, J. E. LeDoux, P. J. Shiromani, T. M. Keane, & J. E. LeDoux (Eds.), *Post-traumatic stress disorder: Basic science and clinical practice* (pp. 375–399). Totowa, NJ: Humana Press.

Rizzo, A. A., Buckwalter, J. G., Bowerly, T. T., van der Zaag, C. C., Humphrey, L. L., Neumann, U. U., &… Sisemore, D. D. (2000). The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, *3*, 483–499.

Rothbaum, B., & Hodges, L. F. (1999). The use of virtual reality exposure in the treatment of anxiety disorders. *Behavior Modification*, *23*, 507.

Rothbaum, B. O., Anderson, P., Zimand, E., Hodges, L., Lang, D., & Wilson, J. (2006). Virtual reality exposure therapy and standard (in vivo) exposure therapy in the treatment of fear of flying. *Behavior Therapy*, *37*, 80–90.

Rothbaum, B. O., Hodges, L., Anderson, P. L., Price, L., & Smith, S. (2002). Twelve-month follow-up of virtual reality and standard exposure therapies for the fear of flying. *Journal of Consulting and Clinical Psychology*, *70*, 428–432.

Rothbaum, B. O., Hodges, L., Smith, S., Lee, J. H., & Price, L. (2000). A controlled study of virtual reality exposure therapy for the fear of flying. *Journal of Consulting and Clinical Psychology*, *68*, 1020–1026.

Rothbaum, B. O., Hodges, L. F., Kooper, R., Opdyke, D., Williford, J. S., & North, M. (1995). Effectiveness of computer-generated (virtual reality) graded exposure in the treatment of acrophobia. *American Journal of Psychiatry*, *152*, 626–628.

Roy, S. S., Klinger, E. E., Légeron, P. P., Lauer, F. F., Chemin, I. I., & Nugues, P. P. (2003). Definition of a VR-based protocol to treat social phobia. *CyberPsychology & Behavior*, *6*, 411–420.

Ruscio, A. M., Brown, T. A., Chiu, W. T., Sareen, J. J., Stein, M. B., & Kessler, R. C. (2008). Social fears and social phobia in the USA: Results from the National Comorbidity Survey Replication. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences*, *38*, 15–28.

Schuemie, M. J., van der Straaten, P., Krijn, M., & van der Mast, C. G. (2001). Research on presence in virtual reality: A survey. *CyberPsychology & Behavior*, *4*, 183–201.

Segal, R., Bhatia, M., & Drapeau, M. (2011). Therapists' perception of benefits and costs of using virtual reality treatments. *CyberPsychology, Behavior & Social Networking*, *14*, 29–34.

Slater, M., Pertaub, D., Barker, C., & Clark, D. M. (2006). An experimental study on fear of public speaking using a virtual environment. *CyberPsychology & Behavior*, *9*, 627–633.

Slater, M., Pertaub, D., & Steed, A. (1999). Public speaking in virtual reality: Facing an audience of avatars. *IEEE Computer Graphics and Applications*, *19*, 6–9.

Smith, T. C., Ryan, M. A. K., Wingard, D. L., et al. (2008). New onset and persistent symptoms of post-traumatic stress disorder self reported after deployment and combat exposures: Prospective population based US military cohort study. *British Medical Journal*, *336*, 366–371.

Solomon, S. D., & Davidson, J. R. T. (1997). Trauma: Prevalence, impairment, service use, and cost. *Journal of Clinical Psychiatry*, *58*, 5–11.

Standen, P. J., & Brown, D. J. (2005). Virtual reality in the rehabilitation of people with intellectual disabilities: Review. *CyberPsychology & Behavior*, *8*, 272–282.

Stendal, K., Balandin, S., & Molka-Danielsen, J. (2011). Virtual worlds: A new opportunity for people with lifelong disability? *Journal of Intellectual and Developmental Disability*, *36*, 80–83.

Tanielian, T., & Jaycox, L. H. (2008). *Invisible wounds of war: Psychosocial and cognitive injuries, their consequences, and services to assist recovery*. Santa Monica, CA: RAND Corporation.

Tortella-Feliu, M., Botella, C., Llabrés, J., Bretón-López, J., del Amo, A., Baños, R. M., & Gelabert, J. M. (2011). Virtual reality versus computer-aided exposure treatments for fear of flying. *Behavior Modification*, *35*, 3–30.

Traylor, A. C., Bordnick, P. S., & Carter, B. L. (2008). Assessing craving in young adult smokers using virtual reality. *American Journal on Addictions*, *17*, 436–440.

Turner, S. M., Beidel, D. C., & Townsley, R. M. (1992). Social phobia: A comparison of specific and generalized subtypes and avoidant personality disorder. *Journal of Abnormal Psychology*, *101*, 326–331.

Villani, D., Riva, F., & Riva, G. (2007). New technologies for relaxation: The role of presence. *International Journal of Stress Management*, *14*, 260–274.

Vincelli, F., Anolli, L., Bouchard, S., Wiederhold, B. K., Zurloni, V., & Riva, G. (2003). Experiential cognitive therapy in the treatment of panic disorders with agoraphobia: A controlled study. *CyberPsychology & Behavior*, *6*, 321–328.

Wald, J., & Taylor, S. (2003). Preliminary research on the efficacy of virtual reality exposure therapy to treat driving phobia. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *6*, 459–465.

Wallace, S., Parsons, S., Westbury, A., White, K., White, K., & Bailey, A. (2010). Sense of presence and atypical social judgments in immersive virtual environments: Responses of adolescents with autism spectrum disorders. *Autism*, *14*, 199–213.

Wallach, H. S., Safir, M. P., & Bar-Zvi, M. (2009). Virtual reality cognitive behavior therapy for public speaking anxiety: A randomized clinical trial. *Behavior Modification*, *33*, 314–338.

Wang, M., & Reid, D. (2009). The virtual reality-cognitive rehabilitation (VR-CR) approach for children with autism. *Journal of CyberTherapy and Rehabilitation*, *2*, 95–104.

Wiederhold, B. K., & Wiederhold, M. D. (2005). A brief history of virtual reality technology. In B. K. Wiederhold & M. D. Wiederhold (Eds.), *Virtual reality therapy for anxiety disorders: Advances in evaluation and treatment* (pp. 11–27). Washington, DC: American Psychological Association.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators & Virtual Environments*, *7*, 225–240.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA: Stanford University Press.

Wood, D., Murphy, J., McLay, R., Koffman, R., Spira, J., Obrecht, R. E., Pyne, J., & Wiederhold, B. K. (2009). Cost effectiveness of virtual reality graded exposure therapy with physiological monitoring for the treatment of combat related post traumatic stress disorder. *Annual Review of CyberTherapy and Telemedicine*, *7*, 223–229.

Wood, P. D., Murphy, J., Center, K., McKay, R., Reeves, D., Pyne, J., Shilling, R., & Wiederhold, B. K. (2007). Combat-related post-traumatic stress disorder: A case report using virtual reality exposure therapy with physiological monitoring. *CyberPsychology & Behavior*, *10*, 309–315.

PART 2

Measurement Strategies for Clinical Psychology

This page intentionally left blank

CHAPTER

7

Assessment and Measurement of Change Considerations in Psychotherapy Research

Randall T. Salekin, Matthew A. Jarrett, *and* Elizabeth W. Adams

Abstract

This chapter focuses on a range of measurement issues including traditional test development procedures and model testing. Themes of the chapter include the importance of integrating research disciplines and nosologies with regard to test development, the need for test development and measurement in the context of idiographic and nomothetic treatment designs, the need for change sensitive measures, and the possible integration of both idiographic and nomothetic treatment designs. Finally, the chapter emphasizes the importance of exploring novel areas of measurement (e.g., biological functioning, contextual factors) using emerging technologies. It is thought that innovative test development and use will lead to improved intervention model testing.

Key Words: assessment, measurement, psychotherapy, treatment, change

As scientists investigate the outcomes of various methods of psychological therapy, they consider a number of scientific issues, such as idiographic versus nomothetic measurement, objective versus subjective assessment, trait versus behavioral change, and the economics of large-scale randomized controlled trials (RCTs; see Chapter 4 in this volume) versus single-subject designs (Barlow & Nock, 2009; Borckardt et al., 2008; Meier, 2008; see Chapter 3 in this volume). Large-scale RCTs have helped to answer questions of aggregate change (e.g., an active treatment showing greater gains than a control condition), but RCTs have yet to fully deliver with regard to individual responses to treatment and the mechanisms of change in psychological therapy (Borckardt et al., 2008). Although moderation analyses in large RCTs address key questions related to individual differences in treatment response, iatrogenic effects of therapy at the individual level may go undetected in RCTs (Barlow, 2010; Bergin, 1966). While this debate ranges from methodological to economic concerns, one critical but often neglected issue is the precision of measurement in capturing psychotherapy change. Specifically, regardless of trial design, how should we measure patient change when interventions are implemented?

There are six primary goals for the current chapter. First, we discuss general issues in test development. Second, we discuss the need for future tests to be able to span current and future nosologies. This issue may become particularly important as we attempt to integrate information that has been garnered across sub-disciplines and fields of research. Third, we discuss special considerations for test development in the context of intervention research. Within this section, we discuss the need to measure not only static but also dynamic characteristics of personality and psychopathology. Fourth, we outline methods to garner more in-depth information on clients during the therapeutic process. Fifth, we provide an overview of new assessment technologies. Sixth, we discuss the important and controversial issues of arbitrary metrics and idiographic versus nomothetic research. It is our hope that a discussion

of these topics will lead to improved assessment technology and an enhanced consideration of measurement issues when evaluating change in psychological therapy.

Test Construction: The Traditional Approach

Many scholarly papers have focused on test construction and assessment (e.g., Clark & Watson, 1995; Salekin, 2009). Historically, test construction was initiated for placement purposes. For example, initial test development focused on selection problems in the early twentieth century (e.g., school and military personnel selection). Tests were designed to be highly efficient in obtaining information about large groups of people in a short amount of time and valuable in terms of predictive validity. In essence, a great emphasis was placed on uniformity and the stable, enduring nature of traits. For instance, Binet used intelligence tests to determine which children would need special education. Alpha I and II tests were developed for World War I and II to determine appropriate settings for soldiers. Test questions helped distinguish those who would be better or less suited for high-risk placements (e.g., fighter pilots; kitchen staff, marines, generals; strategists; mechanics).

This traditional approach to test construction and evaluation has served as the foundation for modern clinical assessment. For example, clinical assessment tools are often evaluated for their psychometric properties, such as scale homogeneity, inter-rater and test–retest reliability, item-response functioning (see Chapter 18 in this volume), and various aspects of validity, including predictive, construct, and ecological validity. The key to successful test development is that the instrument be methodologically sound and have practical implications and utility. Advice for test construction and validity comes from classic works by Cronbach and Meehl (1955), Loevinger (1957), and more recently Clark and Watson (1995). Cronbach and Meehl (1955) argued that investigating the construct validity of a measure entails, at minimum, the following steps: (a) articulating a set of theoretical concepts and their interrelations, (b) developing ways to index the hypothetical concepts proposed by a theory, and (c) empirically testing the hypothesized relations among constructs and their observable manifestations. This means that without a specified theory (the nomological net), there is no construct validity. It has been argued that such articulation of theory is chief in the development of tests. Moreover, a series of studies are needed to determine the construct validity of any measure. Rather than an isolated set of observations, studies must be conducted over time to examine the factor structure, links to other measures, differentiation between selected groups, and hypothesized changes over time or in response to a manipulation. It has been argued that these construct validity studies are critical, in that they guide us toward the most precise instruments.

Loevinger's (1957) monograph continues to be one of the most thorough delineations of theoretically based psychological test construction. Clark and Watson (1995) elaborated on Loevinger's monograph, suggesting that there are a number of important steps in test development and validation. First, test developers should consider whether a test is needed before developing a new test. If new development is unnecessary, then researchers may want to devote time to other knowledge gaps in the literature. There are many reasons for new tests, such as the old ones do not work in the capacity for which they were designed, they do not work as well as they should, a test is required for a different population, there is a need for speedier assessments of a construct (briefer indices), and so on. Should it be decided that a new test is, in fact, needed, Loevinger (1957) and Clark and Watson (1995) provide excellent suggestions for item pool creation (i.e., generate a broad set of items), distillation (i.e., reduce overlap by having experts rate similarity and prototypicality of items), and the subsequent validation process (i.e., linking the construct to meaningful external correlates). These topics have been covered in great length in past research, so we would point readers to these important papers before engaging in test development.

This traditional approach has proven valuable in driving test development. However, we see important next steps in test development to include the need to (a) develop measures that span sub-disciplines and classification systems, (b) include those measures that are appropriate for interdisciplinary research in studies that utilize traditional measures, and (c) design and evaluate measures that are ideal for measuring change in psychotherapy.

The Need for Integration Across Nosologies and Research

The fields of psychiatry and psychology are in the midst of many changes. With the advancement of the DSM-5 and the ICD-11, there are expected changes in the criteria that underpin disorders. The DSM-5 workgroups are also grappling with taxonomic versus dimensional models, as well as the possibility of reducing the overall number of

categories for mental illness. The potential changes to these diagnostic manuals could be vast, and measurement development may be needed with increasing urgency to address DSM and ICD revisions. In addition, there is a clear need for integration of measures across disciplines. For example, with the emergence of fields such as developmental psychopathology (Cicchetti, 1984), a growing body of research linking normal development and abnormal development has emerged (Eisenberg et al., 2005; Rothbart, 2004). We speculate that common tests or constructs that can be used to unite different mental health languages or discipline-specific constructs may be highly valuable. Thus, a new measure might provide a number of labels for similar sets of symptoms (DSM-5 ADHD, ICD-11 Hyperactivity, and FFM Extreme Extraversion).[1] See Table 7.1 for a variety of classification schemes and disorders that likely pertain to similar underlying problems in children.

Integration may not be far off for the mental health fields. Nigg (2006), for example, delineated some of the similarities between classical and temperament theories, highlighting the large degree of overlap of symptoms across systems and the need for integration. At a basic level, scholars have recognized the need to bring together the fields of research (Frick, 2004; Frick & Morris, 2004; Nigg, 2006) to eventually further clinical practice. To illustrate this point, a recent example of cross-fertilization is the research on personality and its potential relation to psychopathology (see Caspi & Shiner, 2006; Nigg, 2006; Rutter, 1987; Salekin & Averett, 2008; Tackett, 2006, 2010; Widiger & Trull, 2007). This research shows connections between personality and psychopathology. Research in this area has underscored three basic human characteristics—approach (extraversion or positive affect), avoidance (fear, or anxiety, withdrawal), and constraint (control)—with hierarchical characteristics that add to these three basic building blocks of human functioning and personality (Watson, Kotov, & Gamez, 2006). Depending on the researcher, some have begun to suggest that a disruption (or an extreme level) in one or more of these areas can reflect psychopathology. For instance, primary deficits in emotion regulation may account for a host of disorders including generalized anxiety, depression, and antisocial conduct (Allen, McHugh, & Barlow, 2011; Brown & Barlow, 2009).

We mention cross-fertilization because we believe that the next step for assessment technology may be to work toward a common cross-discipline nosology that serves to advance knowledge across levels of functioning. One example of such a nosology currently in development within the disciplines of clinical psychology and psychiatry is the initiative by the National Institute of Mental Health (NIMH) to develop what have been called Research Domain Criteria (RDoc). This system is designed to reflect new ways of classifying psychopathology based on dimensions of observable behavior and neurobiological measures. In turn, such a system is working to define basic dimensions of functioning (e.g., fear) that can be studied across levels of analysis (e.g., genes, neural circuits, brain regions, brain functions, etc.). See Table 7.2 for a current draft of the RDoc matrix and examples of genes and brain

Table 7.1 Connecting the Classification Schemes—Childhood Disorders as an Example

Externalizing (Approach) (Positive)	**Internalizing (Withdrawal) (Negative)**
Big 5 Language	
High Extraversion	High Conscientiousness
Low Conscientiousness	High Neuroticism
Low or High Agreeableness	
DSM-IV Classifications	
ODD	Separation Anxiety
CD	Depression
ADHD	Generalized Anxiety
Substance Abuse	Obsessive-Compulsive Disorder
	Eating Disorders
ICD-10	
Hyperkinetic Disorders	Separation Anxiety
Conduct Disorders	Phobic Anxiety
Oppositional Defiant Disorder	Social Anxiety
Temperament	
Surgency	Shyness/Fear
Positive affect/Approach	Irritability/Frustration
Low Control	High Control

Table 7.2 RDoc Matrix Examples

	Units of Analysis							
Domains/ Constructs	**Genes**	**Molecules**	**Cells**	**Circuits**	**Physiology**	**Behavior**	**Self-Reports**	**Paradigms**
Negative Valence Systems								
Active threat ("fear")				Pre-frontal-amygdala circuit	Heart rate			
Potential threat ("anxiety")								
Sustained threat								
Loss								
Frustrative nonreward								
Positive Valence Systems								
Approach motivation								
Initial responsiveness to reward								
Sustained responsiveness to reward								
Reward learning								
Habit								
Cognitive Systems								
Attention	DRD4					Continuous Performance Tests		
Perception								
Working memory							Behavior Rating Inventory of Executive Functioning	
Declarative memory								
Language behavior								
Cognitive (effortful) control								

(*continued*)

Table 7.2 (Continued)

	Units of Analysis							
Domains/ Constructs	**Genes**	**Molecules**	**Cells**	**Circuits**	**Physiology**	**Behavior**	**Self-Reports**	**Paradigms**
Systems for Social Processes								
Imitation, theory of mind								
Social dominance								
Facial expression identification								
Attachment/separation fear								
Self-representation areas								
Arousal/Regulatory Systems								
Arousal & regulation (multiple)								
Resting state activity								

regions that might be implicated for certain conditions (e.g., fear).

Clearly, the integration of biology into psychology and psychological assessment is likely to be important in the upcoming decades. Although the RDoc system is still in development, such multilevel analyses are under way in many research areas. Because of the importance of biology in the assessment process, we mention two areas of study where assessment and measurement of change will be key in the future—genetics and neuroimaging. Because of their rising importance, increasing the precision of assessment will be a priority for further understanding psychopathology.

Measurement in Genetic Studies

Recent articles in *Time* magazine highlight the extent to which genes affect our lives (Cloud, 2010). Behavior and molecular genetics have greatly advanced our knowledge of psychopathology. The field of epigenetics has demonstrated that DNA is not one's destiny—that is, the environment can help change the expression of one's genetic code. Despite significant advances in molecular and behavior genetics, there continues to be a need to integrate genetic research with psychological assessment. Much work will be needed over the next few decades to advance knowledge, and psychology certainly has an important role in assessing psychopathology and its potential alterations over the course of treatment. Specifically, psychology can help by developing precise measures that get us closer to the underlying construct (e.g., disorders) so that genotype–phenotype relationships can be more easily detected. For example, emerging research on the predominantly inattentive type of attention-deficit/hyperactivity disorder (ADHD) has revealed that within-subtype differences may exist with respect to hyperactivity/impulsivity (H/I) symptom count (Carr, Henderson, & Nigg, 2010). Genetic studies and neuropsychological studies are showing support for differentiating between children with ADHD-I who have less than or greater than two H/I symptoms (Volk, Todorov, Hay, & Todd, 2009). In addition, DSM-5 is currently exploring an inattentive presentation (restrictive) subtype (i.e., less than two H/I symptoms). This example relates to the above-mentioned points regarding the integration of genetic research and precise measurement in better understanding genetically driven effects. In addition, better measurement

of environmental factors, including interventions, will be chief if we are to understand what parts of the environment allow for the activation and deactivation of genetic influences.

fMRI Research, Task Development, and Further Psychological Measurement

With the advent of functional magnetic resonance imaging (fMRI), we have gained considerable knowledge regarding the workings of the brain (see Chapter 10 in this volume). However, our knowledge can be furthered through the use of psychological assessment and intervention studies. Because neuroimaging findings are so heavily dependent on specified tasks, advancement in assessment measures is very much needed to shorten the gap between the task performed and the inference made about the participant. Thus, the psychological assessment enterprise could, in part, serve to integrate systems by further developing tasks relevant to imaging studies. Assessing state effects (e.g., affect, thoughts) during neuroimaging tasks could also be beneficial for future research. Overall, interconnecting psychological assessment with cognitive and biological paradigms will help to fasten the disciplines and make more substantial contributions to both basic and applied research.

The above underscores the need for multidisciplinary research but also the need for multimethod–multitrait matrices of behavioral and biological measures. This approach might be one important way to start to integrate information on psychological assessment and to examine progress across treatment. As mentioned, Nigg (2006) and others (Sanislow, Pine et al., 2010) contend that higher-order traits can be conceived as part of a hierarchical model rooted at an abstract level. In contrast, reactivity and basic approach and withdrawal systems may emerge early in life but differentiate into additional meaningful lower-order behavioral response systems during childhood. Nigg (2006) has argued that differentiations can be made at four levels: (a) a frontal limbic, dopaminergically modulated approach system anchored in the reactivity of nucleus accumbens and associated structures to potential reward; (b) a frontal limbic, withdrawal system anchored in reactivity to amygdala and associated structures but also including stress response systems and sympathetic autonomic response, with important distinctions between anxiety and fear response; (c) a constraint system that has diverse neurobiological conceptions but has been conceived as rooted in frontal-striatal neural loops that are dopaminergically modulated and reflect parasympathetic autonomic responsivity as well as the influence of serotonergically modulated reactive control systems; and (d) a related affiliation/empathy system linked to effortful control and the capacity for negative affect. This system subsequently leads to empathy and a desire for and tendency toward affiliation and cooperation (as opposed to social dominance or social interaction, which is driven by the reward/approach systems).

In sum, the uniting of nosologies as well as the integration of biology and psychology will not only serve to provide a common language for understanding "syndromes" or conditions but may also elucidate the processes and primary neural anchors that might be related to a particular condition. In turn, these processes could then become the target of intervention. Next, we turn our attention from single-time-point assessment to considering assessment over multiple time points, and in particular in response to treatment.

Measuring the Benefits of Psychotherapy

Smith and Glass (1977) sparked interest in scientists to quantitatively index change in patients with a standard metric. However, the study of human behavior involves significant sources of error and/or variability, a problem that has affected the precision of measurement as it relates to the prediction of behavior, the definition of psychological constructs (e.g., traits), and capturing change in response to psychological intervention. Measurement issues will always be a challenge when studying a phenomenon as complex as human behavior.

In a review of the history of the assessment process, Meier (2008) argues, much like Loevinger (1957) and Clark and Watson (1995), that when considering test usage, there are three main factors to consider: (1) test content, (2) test purpose, and (3) test context. Significant progress has been made in the area of test content, as mentioned above, but test purpose and test context have been relatively less studied. Given that one purpose of assessment is to measure response to intervention, it seems that measurement development in this area might also be needed. Specifically, in terms of Loevinger (1957) and Cronbach and Meehl's (1955) notion, a theory is required to adequately develop psychological tests that are sensitive to change. A theory of dysfunction and a theory of alleviating the dysfunction may serve as a model that can guide test development in this regard. Also, unlike methods for developing measures that often focus on characteristics or concepts that exhibit stability (e.g., intelligence), a

primary concern for the intervention scientist might be the consideration of measures that adequately tap dynamic characteristics as well as mechanisms of change for symptoms and traits. We discuss these measurement issues next.

Measurement of Psychotherapy Change: Starting with a Theoretical Base

When researchers attempt to measure psychotherapeutic change, they require a theory about how the change occurs. For these purposes, test developers search existing theoretical and empirical literatures to develop items responsive to the interventions and empirical populations in question. The first step would be to conduct a literature review in preparation for creating items and tasks about problems thought to be affected by different interventions with the clinical population of interest. Problems with depression and/or anxiety might be guided by research on coping, courage, or hope. Like the research on general test construction, a thorough explication of the constructs expected to be influenced and those unaffected by the intervention should result in a large initial item pool, providing the capacity to investigate general and specific intervention effects. Test developers might consider including items from theories that conceptualize the process and outcomes of an intervention in different ways as well as items not expected to be affected by treatment. Demonstration of change in intervention-sensitive items and stability in items not expected to change would be another strong method for demonstrating construct validity for change-sensitive tests. It is possible that there are both static and dynamic items, some of which are more appropriate for intervention research (Cook & Campbell, 1979).

Like Loevinger (1957), we suggest that theory should guide the process. For example, Rogers' Client-Centered Psychotherapy provides a nomothetic theory—all individuals are good in nature—so that aspect of human functioning would not be expected to change. However, facilitating one's awareness regarding one's innate good nature and elucidating one's ability to make one's own decisions more confidently could be related to healthy human outcomes and an awareness of one's innate goodness. As such, measures of change might examine level of awareness gained and decision-making frequency and capacity across therapy. Assessment might also then focus on resultant increases in mental health, knowing, at the same time, that one's level of innate goodness (as rated by other Rogerian clinicians) would be stable across that same period. One might also examine a broad network of changes, including phenotypic change on a measure of personality as well as biological changes that might be mapped through genetic expression, electroencephalography (EEG), or fMRI measurement.

Bandura's (1977) self-efficacy theory provides an additional example of a nomothetic theoretical approach that affords specificity regarding what outcomes should change as a result of an intervention. With this theory, outcome expectations refer to knowledge about what specific behaviors will lead to desired outcomes. Self-efficacy expectations, however, refer to beliefs individuals hold about whether they can perform those specific tasks. Anxious individuals may know, for example, that the more they engage in social situations the better they will perform at completing this behavior even if they continue to feel anxious. Thus, they may come to realize that they have competency in completing the task. Bandura (1977) intended this theory to be applicable to all clients and problems and proposed a set of interventions (i.e., modeling, direct performance of behaviors) specifically designed to improve efficacy expectations (see Bandura, 1997). Thus, the model provides an efficient means of measuring mechanisms of change because of the strong relationship to both affect and behavior, and the requisite constructs can be measured through questionnaires or interviews. Biological changes might also be expected and indexed, including approach behavior, which could be detected on the left hemisphere through EEG or imaging work, and increase in heart rate and other biological indicators that show greater approach and motivation.

Although the vast majority of theories for psychotherapies are nomothetic, Mumma (2004) offered a manner in which to conceptualize cases idiographically and also test a case-specific theoretical model from the perspective of the client. He refers to this as "cognitive case formulation" (CCS), where thoughts and beliefs are elicited in using a methodology focusing on clinically relevant issues and situations (Beck, 1995; Persons, 1989; Persons et al., 2001). Once cognitions are elicited, they are examined and selected for relevance to the problem. Following the case formulation, Mumma suggested that the intervention be geared toward the specific cognitions that mediate the problem. Convergent and discriminant validity analyses are then utilized to determine whether the intervention is effective. This approach entails learning about the client's specific cognitions, which may not have any connection to efficacy expectation (as noted in the

nomothetic example above) or the awareness of the patient, but instead to the specific beliefs of the individual, which could then be altered presumably through the course of a tailored psychotherapy. In the section below, we discuss these issues in more detail with regard to nomothetic and idiographic assessment strategies that focus on tapping change.

Assessing Change on Relevant Outcome Constructs in Intervention

As noted throughout the chapter, many assessment measures have been evaluated using traditional psychometric approaches. For example, many of these approaches have come from a trait-based tradition that assumes that specific traits should be static over time. This point is critical, since if trait-based measures are not expected to change, use of such measures in treatment-outcome studies may be inappropriate (Meier, 2008). Traditionally, these measures have included lengthy assessment tools such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1941) and the Child Behavioral Checklist (CBCL; Achenbach, 2001), measures that were developed primarily for selection or screening purposes rather than for assessing change in response to treatment. At the same time, these limitations have been recognized by test developers. For example, a new version of the CBCL was just released that is designed for detecting change (i.e., the BRIEF Problem Monitor). This brief version that utilizes CBCL items was developed for researchers and clinicians interested in examining response to intervention. Similarly, Butcher developed the Butcher Treatment Planning Inventory (Butcher, 1998) as a means of gathering information about a client before treatment as well as to monitor his or her response to treatment. Recently, the Risk-Sophistication-Treatment Inventory has been found to be change sensitive with adolescent offenders (Salekin, Tippey, & Allen, 2012).

In addition to the nature of the assessment tool, informant effects are also present in relation to assessing change. For example, Lambert (1994) found that (a) measures of specific therapy targets produced larger effects than more distal measures, (b) therapist and coder ratings produced larger effects than self-reports, and (c) global ratings produced larger effects than specific symptom ratings. Overall, these findings point toward the idea that there are "reliable differences in the sensitivity of instruments to change" (Lambert, 1994, p. 85). In addition to the issue of measurement, one must also consider timing and changes prior to intervention. For example, Kelly, Roberts, and Ciesla (2005) found that clients exhibited significant change in symptoms of depression prior to the start of intervention. In some instances, studies that conduct an initial assessment and then begin treatment weeks later may miss an important window of assessing change (i.e., changes between initial assessment and the start of treatment).

If change-sensitive measures are seen as a priority, then it will be necessary to discuss how such measures might be developed. Prior to discussing recommendations for assessing the psychometric properties of change-sensitive measures, we briefly discuss recommendations that have been identified for item selection for change-sensitive measures. The process for developing such measures is similar to what Cronbach and Meehl (1957) and Loevinger (1957) have suggested for trait measures. Meier (2008) recommends a series of intervention item-selection rules that align well with traditional theories regarding test development (see Table 7.3).

As with traditional tests, items created for change-sensitive tests should be theoretically based. Contemporary measurement theories purport that test validation is closely related to traditional hypothesis testing (Ellis & Blustein, 1991; Landy, 1986). Theoretical grounding also provides an important context for understanding the meaning of changing values of items in a change-sensitive assessment.

In relation to the assessment of reliability and validity, the general recommendations discussed in

Table 7.3 Intervention Item Selection Rules

1. Ground scale items in theoretical and empirical literature relevant to applicable interventions, clinical populations, and target problems.
2. Aggregate items at appropriate levels, and assess range of item scores at pretest.
3. Make sure items evidence change in intervention conditions (in theoretically expected direction).
4. Examine whether differences in change exist between intervention and comparison groups.
5. Examine whether intake differences exist between comparison groups.
6. Examine relations between item scores and systematic error sources.
7. Cross-validate results to minimize chance effects.

Based on Cronbach and Meehl (1955), Loevinger (1957), and recently Meier (2008).

the traditional test-development section apply with some special considerations. It is still important to assess for the stability of a construct in the absence of intervention. As Vermeersch and colleagues (2004, p. 38) noted, "the sensitivity to change of a measure is directly related to the construct validity of the instrument." Once basic psychometric properties are established (e.g., internal consistency, test–retest reliability, inter-rater reliability, etc.), one can then evaluate how sensitive to change the measure is in response to treatment. Meier (2008) noted that once an outcome measure is established as a stable measure with strong psychometric properties, it should not change in response to repeated administration in the absence of intervention but should change in the presence of intervention. In turn, such an evaluation would require the examination of test–retest reliability in a control sample, and the test–retest range should approximate the periods used between outcome measures in the clinical sample

Table 7.4 Cross-Cutting Areas of Change Across Diverse Psychotherapies: Potential Change-Sensitive Items

Domain	Description
Thought	
Listens	Pays attention to others and follows through with instructions
Shares thinking	Is able to talk about what he or she is thinking about with others
Rational	Is able to develop rational thoughts with a number of options for problems
Views of the self	Self-understanding
Mood	
Bright affect	Becomes happier with self over the course of treatment
Emotion regulation	Is able to regulate emotions, control temper
Compensatory skills	Is able to engage in behaviors to alleviate or reduce negative feelings
Behavior	
Helping	Contributes in the home helping with chores and in the community
Sports/Hobbies	Engages in sports and/or hobbies
Well behaved	Behaved at home and at school/work
Ways of responding	Learns ways of responding
Interpersonal	
Close friend	New friends or a close friend attained. A close or confiding high-quality friendship. Is able to start conversations.
Social life	Number of friends, frequency of social activities, level of conflict with others; makes friends easily
Romantic	Quality and satisfaction with romantic relationship (adolescents and young adults)
Family	Quality of relationship with parents and siblings
Non-Interpersonal	
Academic	Quality of school performance; level of effort and grades
Finance	Financial standing. How does the individual handle money? (adolescents and young adults)
Personal health	Health status and healthy lifestyle
Family health	Health status and healthy lifestyle in immediate family

(e.g., weekly, monthly, etc.). Table 7.4 provides an example of several potential change-sensitive items that could be included in test–retest designs.

Changes are to be expected in biology as well. For instance, as we chart out differences based on extraversion and introversion or approach and withdrawal, we may expect certain changes in biology, such as a set of neurons that becomes more active in the amygdala. However, more subtle differences may also be expected, and measurement would be needed to take into account such differences. Specifically, although we labeled our two broad categories as approach or withdrawal, biological distinctions are not always similarly demarcated and grouped. Recent physiological evidence is consistent with this perspective. For example, left frontal activation appears to be related to motivation (approach) more clearly than to valence. Those measuring change across intervention trials would need to be aware of this. Moreover, Nigg (2006) has argued that angry affect sometimes reflects a potential to overcome obstacles in anticipation of reward or reinforcement emanating from traits such as surgency and reward responsivity, whereas in other instances it reflects a propensity to resentful, angry, hostile affect that is responsive to fear or anticipated negative consequences. These may be related to panic, alarm, and/or stress response, as well as rage related to panic or fear. As mentioned, we expect that biological research in conjunction with psychological assessment will advance the field in this regard. However, for now, they are important factors to be cognizant about in designing measures and indices of change.

Specific Considerations for Treatment Outcome Evaluation Methodologies

It has been argued that current assessment and test-development practices may pose some problems in the area of treatment research or may not capture all the important variables when considering whether or not an individual has changed. In addition to assessment challenges, there are controversies in relation to experimental methodologies that are best used for evaluating treatment. This seems like a good point in the chapter to discuss concerns regarding methodology because these concerns are also linked to a general uneasiness about the types of assessments that are needed to detect change in clients. With the movement toward establishing empirically supported treatments (Kendall, 1998), there has been an increasing emphasis on the use of the RCT, a method that allows randomization to treatment or control conditions. Such a movement has coincided with the general trend within the field of medicine to establish guidelines for evidence-based medicine. Strengths of the RCT include enhanced causal clarity and the ability to utilize parametric statistics to examine treatment outcomes (see also Chapter 4 in this volume). At the same time, RCTs have limitations (e.g., Barlow & Nock; 2009; Borckardt et al., 2008; Westen, Novotny, & Thompson-Brenner, 2004). Although RCTs can identify significant differences at the group or aggregate level, clients often show substantial variability in response to treatment, and the methodology of the RCT is less well equipped to address this variability (Barlow, 2010). More importantly, although many trials have supported the efficacy of therapeutic approaches for a range of psychological problems, there is still a dearth of knowledge regarding the mechanisms of change in treatment. These issues have long challenged clinicians and researchers. For example, Kiesler (1966) first posed the question of personalized treatment approaches. Although this level of specificity remains a challenge, the question remains whether the current assessment tools and methodologies allow us to move closer to determining what treatments are more effective for certain individuals.

New methodologies may be needed. With any methodology, there will be a need for assessment that is capable of indexing change, a crucial aspect for determining the effects of treatment. Although earlier parts of the chapter have focused on traditional test-development procedures, model testing, and the development of change-sensitive tests, the final part of the chapter will focus on selecting measures in the context of treatment evaluation that might best capture change given the particular strengths of the treatment design. Prior to discussing test selection in the context of treatment, we briefly review the evidence-based treatment (EBT) movement and the treatment evaluation methodologies below that have been supported by this movement.

Empirically Supported Methodologies

Although the movement to evaluate the efficacy of psychosocial treatments occurred many years prior to 1995, the first report on evidence-based practice was issued at that time. This report, issued by the Society of Clinical Psychology Task Force on Promotion and Dissemination of Psychological Procedures, established guidelines for both methodologies and the level of evidence needed for various classifications of "empirical support." Three categories were established for empirically supported treatments: (1) well-established treatments, (2) probably efficacious treatments, and (3) experimental treatments.

The primary distinction between *well-established treatments* and *probably efficacious treatments* is that the former must have been found to be superior to a psychological placebo, pill, or another treatment whereas the latter must prove to be superior only to a waitlist or no-treatment control condition. In addition, the former require evidence from at least two different investigatory teams, whereas the latter require evidence from only one investigatory team. For both types of empirically supported treatments, client characteristics (real clients) should be well specified (e.g., age, sex, ethnicity, diagnosis), and the clinical services should follow written treatment manuals. *Experimental treatments* are those treatments not yet shown to be at least probably efficacious. This category was intended to capture treatments frequently used in clinical practice but not yet fully evaluated *or* newly developed ones not yet put to the test of scientific scrutiny.

Most clinicians and researchers are aware of these guidelines as they relate to group comparisons, but single-case designs were also supported in this task force report. Single-case designs have a long history in psychology, and many single-case studies led to advances in clinical applications. Nevertheless, their use has decreased substantially as the National Institutes of Health funded large-scale RCTs. More recently, single-case designs have re-emerged (Barlow & Nock, 2009; Borckardt et al., 2008; Westen & Bradley, 2005). RCTs provide significant causal clarity through the use of randomization, but they tend to answer questions of aggregate change (e.g., do clients in the treatment condition improve more than those in a control condition?). Several large-scale RCTs have established that certain treatments are more effective than credible placebos or alternative treatments, but some of these studies, particularly those that fail to address treatment mediators and moderators, have struggled to describe (a) idiographic change and (b) processes of change in treatment. In comparison to RCTs, single-case designs may lack the causal clarity of RCTs given the lack of randomization, but at the same time, they are helpful in examining idiographic change (see Chapter 3 in this volume). In the following sections, we will outline measurement selection strategies for both nomothetic or group designs as well as idiographic or single-case designs. In addition, we discuss areas of integration that would allow for maximizing treatment evaluation.

Nomothetic Designs

Nomothetic designs typically assess functioning prior to treatment (e.g., pretreatment, baseline), possibly during the middle of treatment, at the end of treatment (posttreatment), and then again at a follow-up point. Increasingly, RCTs are incorporating weekly assessments of key outcomes, affording opportunities for more sophisticated evaluations of treatment-related changes (e.g., Kendall, Comer, Marker, Creed, Puliafico, et al., 2009). Such designs may incorporate both short-term and longer-term follow-up assessments. These models expect that individuals have the same traits or behaviors to varying degrees, given that all individuals often meet criteria for the same disorder or related disorders. While the nomothetic method is likely to always be useful as the most rigorous method with which to afford causal conclusions, the nomothetic approach has been criticized. First, RCTs often report aggregate data, but idiographic researchers might be more interested in individuals who responded particularly well versus those who did not change. Fortunately, formal evaluation of treatment moderators is becoming increasingly common in RCT analyses (see Chapter 15 in this volume). A second criticism is that in applied settings, clinicians question the applicability of findings from RCTs to the individuals seen in their clinical settings. In other words, despite RCTs using real clients who meet diagnostic criteria, there is a perception that problems exist in generalizing nomothetic results to an idiographic situation. Although we agree with some of the criticisms of RCTs, these designs nevertheless still have great value in treatment evaluation. There may be room to integrate some of the advantages of idiographic designs into future RCTs.

Idiographic Designs

The intensive study of the individual has netted some of psychology's important discoveries. The founders of experimental psychology, including Fechner, Wundt, Ebbinghaus, and Pavlov, all studied individual organisms and produced field changing findings. Allport (1937, 1938) was one of the first to discuss the importance of the idiographic approach with respect to personality. While nomothetic tests gained popularity because of their ease of use and economics, it has been argued that idiographic methods should climb in importance because of their potential for further improving validity, particularly in clinical settings. Although the previously discussed measures can be utilized in both nomothetic and idiographic treatment designs, special consideration is needed when utilizing an idiographic design. Idiographic designs might focus more on the specific problems of the individual,

and measures may be tailored to the intervention process. Testing procedures may also involve direct observation of the trait or behavior. Another consideration is the frequency of measurement. Many idiographic designs may include daily or weekly measurement of symptoms or traits and behavior. In turn, this frequency of measurement has implications for detecting change and the analysis of data.

Although in general frequently assessing a problem can be beneficial to understanding change that is occurring, one common problem when variables are frequently measured over the course of intervention is the problem of autocorrelation. Conventional parametric and nonparametric statistics assume that observations are independent, but data collected in single-case designs are highly dependent. Autocorrelation, sometimes referred to as serial dependence, reflects the fact that one value in time depends on one or more of the preceding values. For example, weather is a natural example of autocorrelation: the weather yesterday is a good predictor of the weather today. In turn, single-case design approaches need statistical analysis tools that take into account factors such as autocorrelation. Recently, statistical approaches have emerged to begin to address such issues (see Borckardt et al., 2008). One of the key assumptions, though, is equal intervals of measurement (e.g., weekly, daily, etc.). In turn, researchers seeking to use single-case design elements need to plan for having evenly spaced assessment intervals. Finally, approaches that use such frequent measurement have the capacity to capture the mechanisms of change in therapy. For example, an RCT that utilizes only a mid-treatment assessment may miss out on important changes in the therapeutic process. In addition, new single-case design approaches also allow for the examination of multivariate process change using cross-lagged correlations (e.g., does variable 1 change before variable 2? If so, what was the lag in change?). Self-monitoring, natural-language, and word-use approaches can all be useful in gleaning further idiographic information about a client (Dimaggio & Semerari, 2004). See Table 7.5 for examples of idiographic methods.

Physiological Data in Natural Settings

Physiological data may be collected in natural settings and more frequently across the week. Such data have been collected with respect to cardiovascular, respiratory, and hormonal systems. In one study examining panic disorder, 12 subjects wore an apparatus to record ECG as well as signals from an event marker for 12 hours a day across a 2-day period (Freedman, Ianni, Ettedgui, & Puthezhath, 1985). A number of other naturalistic studies have used similar techniques with larger samples to examine the relation between cardiovascular functioning and anxiety (e.g., Hoehn-Saric, McLeod, Funderburk, & Kowalski, 2004). Other research has examined respiratory functioning (Klein, 1993; Ley, 1985) or hyperventilation/hypocapnia (i.e., reduced level of carbon dioxide in the blood). Endocrine measurement has also been examined for anxiety disorders since they are thought to be related to psychological stress. Psychological stress in humans leads to a cascade of hormonal changes regulated by the hypothalamic-pituitary-adrenal (HPA) axis, with an increase in cortisol being the most typically observed finding (Alpers, 2009).

Integration of Nomothetic and Idiographic Elements

We have discussed the pros and cons of nomothetic and idiographic designs. The debate has led many researchers to utilize only one of the approaches, yet we see substantial room for integration. For example, researchers conducting an RCT could easily integrate weekly assessment into the framework of the RCT assessment schedule (e.g., short weekly ratings of symptoms in addition to more comprehensive assessments at pretreatment, midtreatment, and posttreatment). As mentioned, some studies now integrate the two. In addition, researchers could continue with brief weekly assessments in the short period between the initial pretreatment assessment and the start of treatment. For example, in many RCTs, additional assessment sessions and/or other administrative processes (e.g., establishing diagnoses, randomizing to a treatment condition, etc.) may occur in the weeks following the pretreatment assessment. Lack of measurement during this period could result in a loss of information regarding whether change occurs in the weeks prior to the start of treatment. As mentioned earlier, weekly assessment during RCTs would also allow for evaluation of session-by-session change in addition to changes between key comprehensive assessment points (e.g., pretreatment, midtreatment, etc.). Finally, this weekly assessment also allows for an examination of individual change. For example, researchers utilizing both approaches would be able to examine aggregate change but also processes of change at the individual level. Although seen in only a handful of reported RCTs, the merits of more regular assessments and the feasibility of their

Table 7.5 Idiographic Outcome Measures

Progress Notes	A therapist's notes provide the natural language of individuals.
Client Notes	Therapists may ask clients to keep a small notebook with them, jotting down notes that seem most applicable to them. They may do this to record incidents and what occurred just before and right after the incident.
Client Diaries	Clients may keep daily diaries of their life, and this may serve as an important key to cognitions that are related to a problem and thus a factor to monitor across the intervention.
Writing Assignments	Patients may participate in writing assignments to allow assessment of their personality. This might involve writing narratives about the way they might solve a current problem they have.
Storytelling	Clients may tell stories about family members and how they interact with those family members.
Self-monitoring	Self-monitoring is a combination of self-report and behavioral observation whereby individuals observe and record behaviors intended to be the targets of an intervention (Kazdin, 1974). Monitoring provides treatment-sensitive data in a wide variety of treatment domains. A client may be assigned to monitor problematic behaviors in preparation for the intervention. Traditionally, the procedure teaches clients to keep a notebook (or electronic device) with them at all times to record behaviors immediately after they occur; only a single response is recorded at a time (Cone, 1999; Barton, Blanchard, & Veazy, 1999).
Natural Language	Verbal and nonverbal communications are the primary media through which therapist and client transfer information and meaning. One method of understanding individuals well is to listen closely to what they say—that is, to study their use of language. A focus on natural language is the essence of numerous narrative approaches to studying, assessing, and intervening with individuals. Two such contemporary approaches to natural language that offer potential insights into improvement in the measurement of psychological constructs are narrative therapy and Pennebaker's word analysis.

integration within an RCT would combine to make for a more potent, and informative, RCT.

Although we have discussed modifications to the RCT, single-case designs can also benefit from nomothetic approaches. For example, while the single-case researcher may be interested in individual-level change, standard pre–post analyses can also be conducted. Although such pre–post analyses often involve a small sample size, nonparametric equivalents to *t* tests and repeated-measures ANOVA, such as Wilcoxon and Friedman tests, can be utilized in such circumstances (see ter Kuile et al., 2009, for an example). Overall, we see substantial room for integrating the strengths of both methodologies.

We have focused on change-sensitive measures which may lead one to believe that we are focusing on easily changeable behaviors, but we do not believe that the field should move away from treating difficult conditions or even traits (see Tang, DeRubeis, Hollon, Amsterdam, Shelton, & Schalet, 2009). That is, when we talk about change-sensitive tests, many of these tests are designed to examine behavior, but we may also be interested in whether we can effectively change broader problems. Thus, it might be helpful to determine if we can alter, for example, depression, psychotic symptoms, or personality traits such as "openness to experience," characteristics that are often thought to be stable. Examination of such changes might be looked at in terms of cognition and behavior. With respect to openness, one might measure the extent to which the client agrees to meet other people, try new restaurants, or see new movies; these activities are open and outside of the norm for the individual. Similarly, there may be changes that are possible with respect to difficult personalities in youth and adults such as interpersonal callousness (see Salekin, 2002; Salekin, Lester, & Sellers, 2012; Salekin et al., 2012; Salekin, Worley, & Grimes, 2010; see also Hawes & Dadds, 2007).

When seeking these change-sensitive measures, it will be important to keep in mind that we want our measures to accurately tap the constructs we intend for them to tap, as well as for the measures to have meaning along the dimension (at high levels of depression, the person is actually depressed) and that what we are calling the measure (e.g., Dep

Scale-Revised) is in fact what the measure is indexing (e.g., self-esteem). Correspondingly, we would want the change to be clinically significant (Jacobson & Traux, 1991; Kendall, 1998) and not reflect issues with measurement. In recent years, there has been some concern about arbitrary metrics. If measures are not sound, it can be very problematic for research on interventions. As Kazdin (2006) puts it, "Arbitrary metrics and related points about them… if applied, shake key pillars of the foundations of EBTs and greatly qualify the conclusions that can be made about these treatments" (p. 45).

A solution to this issue can be found from research advice previously published in the 1970s and 1980s. Specifically, researchers during this time were encouraged to evaluate whether changes in treatment were socially valid; this meant that researchers were asked to focus on domains that were important to individuals in everyday life. Questions were further asked as to whether changes following treatment actually made a difference in clients' lives. Specifically did the clients notice any difference following treatment and did the intervention also make a difference to those with whom the clients' interacted (e.g., relatives and friends; Wolf, 1978). Although there are likely many ways to accomplish this clinical research goal, one method previously used suggests the use of naïve observers who rated the performance of clients prior to and after treatment (Kazdin, 1977). Technology may further assist with this needed component to research studies, where family and friends will be able to provide feedback on issues pertaining to clinical change.

Context and Assessment Technology: Needed Advancements to Keep Pace with Psychotherapy Innovations

Technology is changing the way that clinicians and researchers perceive assessment and treatment with their clients. The prevalent use of computers, personal digital assistants, and mobile phones could very well help researchers examine progress in therapy from nomothetic and idiographic perspectives. Given our need to more closely test the theoretical models of intervention, such devices offer unique methods of measuring behavior, assessing outcomes, and delivering treatments that are easily accessible, thereby, increasing response frequency and accuracy in comparison to paper measures. To underscore this point, measuring individuals in natural settings at specific moments, often referred to as ecological momentary assessment (EMA; see Chapter 11 in this volume), or intensive repeated measures in naturalistic settings (IRM-NS), would allow for the assessment of mood, thoughts, and behaviors in the moment while offering a more comprehensive data collection than measures administered during a therapy session (Heron & Smyth, 2010; Moskowitz, Russell, Sadikaj, & Sutton, 2009; Palermo, 2008; Trull & Ebner-Priemer, 2009; Wenze & Miller, 2010). Also, reminders can be sent to patients about homework assignments, and other cues can be given. While paper-and-pencil tests are likely to continue to serve a purpose, critics have argued that in-therapy self-report measures are subject to recall biases, semantic memory, heuristics, and preexisting beliefs, which these technological advances could reduce (Moskowitz et al., 2009). Technology has advanced and offers hope for better measurement, yet several aspects of using technology will continue to require further investigation for appropriate use, including psychometric properties, method choice, and technological variables.

Overall, assessments that use computers, personal digital assistants, and mobile phones will likely be important as they are tangible and adapt clinical measures to the technological progression of society. Being familiar with these devices and their use in therapy offers clinicians a unique method of measurement. Further, research needs to be conducted to provide guidelines for implementing therapy via technology, but the potential for enhanced clinical utility is present (see Table 7.6). Such research may begin to address some of the issues raised by Mischel (1968), who ignited a controversy in the late 1960s about the extent to which the environment and context had to do with an individual's personality (see also Kantor, 1924; Lewin, 1935; Murray, 1951; Peterson, 1968). When measuring change, a method that combines event-specific reactions and the frequencies of the events themselves could also prove most fruitful. Such integration may converge better with what is assessed by the overall test battery. This may provide additional information on person–situation mixtures requiring careful consideration (e.g., verbal aggression may be charged only

Table 7.6 Technological Advances

Computers
Smart phones
Personal digital assistants
Global positioning systems

by one parent in a certain context) and could help with concerns regarding arbitrary metrics.

Concluding Comments: Integration and Future Directions

Improving the assessment of psychological variables will be an important future task for psychology. This chapter covered a number of important topics such as test construction, integration of nomological systems of classification, the importance of biology, measurement of change, development of new measures, the integration of nomothetic and idiographic designs, and technological advances, as well as a brief discussion on arbitrary metrics. We see value in being more idiographic in our research, and technology is offering us more opportunities to do so. Although most psychological researchers have been trained in group comparison designs and have relied primarily on them, exciting advances have been made in the use of idiographic methodologies, such as the single-case, experimental design (see Barlow, Nock, & Hersen, 2009).

We see value in placing more emphasis on an integration of nomothetic and idiographic strategies that can be used in both clinical and basic science settings. As noted by Barlow and Kazdin, in clinical science, having established the effectiveness of a particular independent variable (e.g., an intervention for a specific form of psychopathology), one could then carry on with more idiographic efforts tracking down sources of inter-subject variability and isolating factors responsible for this variability (see also Kazdin & Nock, 2003; Kendall, 1998). This might allow us to better assess change in psychological therapy. Necessary alterations in the intervention protocols to effectively address client variability could be further tested, once again idiographically, and incorporated into the treatments. Researchers in basic science laboratories could undertake similar strategies and avoid tolerating the large error terms. By incorporating a number of the innovations in the assessment field, psychological science, both basic and applied, could make significant strides in psychological therapy research and practice in the years to come.

Note

1. Aside from integrating information across disciplines, integration of information across sources has also become an important consideration. Although beyond the focus of this chapter, we see the need to better understand and integrate information across sources as another important step for researchers to take in the assessment domain (see De Los Reyes & Kazdin, 2004, 2005).

References

Achenbach, T. M. (2001a). *Manual for the Child Behavior Checklist 4–18 and 2001 Profile*. Burlington: University of Vermont, Department of Psychiatry.

Allen, L. B., McHugh, R. K., & Barlow, D. H. (2008). Emotional disorders: A unified protocol. In D. H. Barlow (Ed.), *Clinical handbook of psychological disorders: A step-by-step treatment manual* (4th ed., pp. 216–249). New York: Guilford Press.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Henry Holt.

Allport, G. W. (1938). Editorial. *Journal of Abnormal and Social Psychology*, *33*, 3–13.

Alpers, G. W. (2009). Ambulatory assessment in panic disorder and specific phobia. *Psychological Assessment*, *21*, 476–485.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review*, *84*, 191–215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist*, *65*, 13–20.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*, 19–21.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single-case experimental designs: Strategies for studying behavior change*. New York: Pearson.

Barton, K. A., Blanchard, E. B., & Veazy, C. (1999). Self-monitoring as an assessment strategy in behavioral medicine. *Psychological Assessment*, *11*, 490–497.

Beck, A. T. (1995). Cognitive therapy: Past, present, and future. In M. J. Mahoney (Ed.), *Cognitive and constructive psychotherapies: Theory, research, and practice* (pp. 29–40). New York: Springer Publishing.

Bergin, A. E. (1966). Some implications of psychotherapy research for therapeutic practice. *Journal of Abnormal Psychology*, *71*, 235–246.

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. *American Psychologist*, *63*, 77–95.

Brown, T. A., & Barlow, D. H. 2009). A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: Implications for assessment and treatment. *Psychological Assessment*, *21*, 256–271.

Butcher, J. N. (1998). *Butcher Treatment Planning Inventory*. San Antonio, TX: The Psychological Corporation.

Carr, L., Henderson, J., & Nigg, J. T. (2010). Cognitive control and attentional selection in adolescents with ADHD versus ADD. *Journal of Clinical Child & Adolescent Psychology*, *39*, 726–740.

Caspi, A., & Shiner, R. L. (2006). Personality development. In W. Damon & R. Lerner (Series Eds.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional and personality development* (6th ed., pp. 300–365). Hoboken, NJ: Wiley.

Cicchetti, D. (1984). The emergence of developmental psychopathology. *Child Development*, *55*, 1–6.

Clark, L. A., & Watson, D. B. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Cloud, J. (2010). Why genes aren't our destiny: The new field of epigenetics is showing how the environment and your

choices can influence your genetic code—and that of your kids. *Time*, January, 48–53.
Cone, J. D. (1999). Introduction to the special section on self-monitoring: A major assessment method in clinical psychology. *Psychological Assessment*, *11*, 411–414.
Cook, T., & Campbell, D. (1979). *Quasi-experimentation*. Chicago, IL: Rand McNally.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin*, *52*, 281–302.
De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment*, *16*, 330–334.
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*, 483–509.
Dimaggio, G., & Semerari, A. (2004). Disorganized narratives. In L. E. Angus & J. McLeod (Eds.), *Handbook of narrative and psychotherapy* (pp. 263–282). Thousand Oaks, CA: Sage.
Eisenberg, N., Sadovsky, A., Spinrad, T. L., Fabes, R. A., Losoya, S. H., Valiente, C.,…Shepard, S. A. (2005). The relations of problem behavior status to children's negative emotionality, effortful control, and impulsivity: Concurrent relations and prediction of change. *Developmental Psychology, 41,* 193–211.
Ellis, M. V., & Blustein, D. L. (1991). Developing and using educational and psychological tests and measures: The unificationist perspective. *Journal of Counseling & Development*, *69*, 550–555.
Freedman, R. R., Ianni, P., Ettedgui, E., & Puthezhath, N. (1985). Ambulatory monitoring of panic disorder. *Archives of General Psychiatry*, *42*, 244–248.
Frick, P. J. (2004). Integrating research on temperament and childhood psychopathology: Its pitfalls and promise. *Journal of Clinical Child and Adolescent Psychology*, *33*, 2–7.
Frick, P. J., & Morris, A. (2004). Temperament and developmental pathways to conduct problems. *Journal of Clinical Child And Adolescent Psychology*, *33*(1), 54–68.
Hathaway, S. R., & McKinley, J. C. (1941). *The Minnesota Multiphasic Personality Inventory manual*. New York: Psychological Corporation.
Hawes, D. J., & Dadds, M. R. (2007). Stability and malleability of callous unemotional traits during treatment for childhood conduct problems. *Journal of Clinical Child and Adolescent Psychology, 36,* 347–355.
Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*, 1–39.
Hoehn-Saric, R., McLeod, D. R., Funderburk, F., & Kowalski, P. (2004). Somatic symptoms and physiologic responses in generalized anxiety disorder and panic disorder. *Archives of General Psychiatry*, *61*, 913–921.
Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.
Kantor, J. R. (1924). *Principles of psychology* (Vol. 1). Bloomington, IN: Principia Press.
Kazdin, A. E. (1974). Self monitoring and behavior change. In M. J. Mahoney & C. E. Thorsen (Eds.), *Self-control: Power to the person* (pp. 218–246). Monterey, CA: Brooks-Cole.
Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1,* 427–452.
Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61,* 42–49.
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.
Kelly, M. A. R., Roberts, J. E., & Ciesla, J. A. (2005). Sudden gains in cognitive behavioral treatment for depression: When do they occur and do they matter? *Behavior Research and Therapy*, *43*, 703–714.
Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology*, *66*, 3–7.
Kendall, P. C., Comer, J. S., Marker, C. D., Creed, T. A., Puliafico, A. C., Hughes, A. A., Martin, E., Suveg, C., & Hudson, J.L. (2009). In-session exposure tasks and therapeutic alliance across the treatment of childhood anxiety disorders. *Journal of Consulting and Clinical Psychology*, *77*, 517–525.
Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, *65*, 100–136.
Klein, D. F. (1993). False suffocation alarms, spontaneous panics, and related conditions: An integrative hypothesis. *Archives of General Psychiatry*, *50*, 306–317.
Lambert, M. J. (1994). Use of psychological tests for outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 75– 97). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.
Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*, 1183–1192.
Lewin, K. (1935). *A dynamic theory of personality*. New York, McGraw-Hill.
Ley, R. (1985). Agoraphobia, the panic attack and the hyperventilation syndrome. *Behaviour Research and Therapy, 23,* 79–81.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.
Meier, S. T. (2008). *Measuring change in counseling and psychotherapy*. New York: Guilford Press.
Mischel, W. (1968). *Personality and assessment*. Hoboken, NJ: John Wiley & Sons Inc.
Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology*, *50*, 131–140.
Mumma, G. H. (2004). Validation of idiosyncratic cognitive schema in cognitive case formulation: An intraindividual idiographic approach. *Psychological Assessment*, *16*, 211–230.
Murray, H. A. (1951). Uses of the Thematic Apperception Test. *The American Journal of Psychiatry, 107,* 577–581.
Nigg, J. T. (2006). Temperament and developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *47*, 395–422.
Palermo, T. M. (2008). Editorial: Section on innovations in technology in measurement, assessment, and intervention. *Journal of Pediatric Psychology*, *33*, 35–38.
Peterson , D. R. (1968). *The clinical study of social behavior*. East Norwalk, CT: Appleton-Century-Crofts.
Persons, J. B. (1989). Cognitive Therapy in Practice: A Case Formulation Approach. New York: W.W. Norton.
Persons, J. B., Davidson, J., & Tompkins, M. A. (2001). *Essential Components of Cognitive-Behavior Therapy for Depression*. Washington: American Psychological Association.

Rothbart, M. K. (2004). Commentary: Differentiated measures of temperament and multiple pathways to childhood disorders. *Journal of Clinical Child and Adolescent Psychology*, *33*, 82–87.

Rutter, M. (1987). Temperament, personality and personality disorder. *British Journal of Psychiatry*, *150*, 443–458.

Salekin, R. T. (2002). Psychopathy and therapeutic pessimism: Clinical lore or clinical reality? *Clinical Psychology Review*, *22*, 79–112.

Salekin, R. T. (2009). Psychopathology and assessment: Contributing knowledge to science and practice. *Journal of Psychopathology and Behavioral Assessment*, *31*, 1–6.

Salekin, R. T., & Averett, C. A. (2008). Personality in childhood and adolescence. In M. Hersen & A. M. Gross (Eds.), *Handbook of clinical psychology (Vol. 2): Children and adolescents* (pp. 351–385). Hoboken, NJ: John Wiley & Sons.

Salekin, R. T., Lester, W. S., & Sellers, M. K. (2012). Mental sets in conduct problem youth with psychopathic features: Entity versus incremental theories of intelligence. *Law and Human Behavior, 36,* 283–292.

Salekin, R. T., Tippey, J. G., & Allen, A. D. (2012). Treatment of conduct problem youth with interpersonal callous traits using mental models: Measurement of risk and change. *Behavioral Sciences and the Law, 30,* 470–486.

Salekin, R. T., Worley, C., & Grimes, R. D. (2010). Treatment of psychopathy: A review and brief introduction to the mental model approach for psychopathy. *Behavioral Sciences and the Law*, *28*, 235–266.

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., Wang, P. S., & Cuthbert, B. N. (2010). Developing constructs for psychopathology research: Research domain criteria. *Journal of Abnormal Psychology*, *119*, 631–639.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Tackett, J. L. (2006). Evaluating models of the personality-psychopathology relationship in children and adolescents. *Clinical Psychology Review*, *26*, 584–599.

Tackett, J. L. (2010). Measurement and assessment of child and adolescent personality pathology: Introduction to the special issue. *Journal of Psychopathology and Behavioral Assessment, 32,* 463–466.

Tang, T. Z., DeRubeis, R. J., Hollon, S. D., Amsterdam, J., Shelton, R., & Schalet, B. (2009). Personality change during depression treatment. *Archives of General Psychiatry*, *66*, 1322–1330.

ter Kuile, M. M., Bulte, I., Weijenborg, P. T., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology*, *77*, 149–159.

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment*, *21*, 457–462.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology*, *51*, 38–49.

Volk, H. E., Todorov, A. A., Hay, D. A., & Todd, R. D. (2009). Simple identification of complex ADHD subtypes using current symptom counts. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*, 441–450.

Watson, D., Kotov, R., & Gamez, W. (2006). Basic dimensions of temperament in relation to personality and psychopathology. In R. F. Krueger & J. L. Tackett (Eds.), Personality and psychopathology (pp. 7–38). New York: Guilford Press.

Wenze, S. J., & Miller, I. W. (2010). Use of ecological momentary assessment in mood disorders research. *Clinical Psychology Review*, *30*, 794–804.

Westen, D., & Bradley, R. (2005). Empirically supported complexity: Rethinking evidence-based practice in psychotherapy. *Current Directions in Psychological Science*, *14*, 266–271.

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, *130*, 631–663.

Widiger, T. A., & Trull, T. J. (2007). Plate tectonics in the classification of personality disorders. *American Psychologist*, *62*, 71–83.

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.

CHAPTER

8 Observational Coding Strategies

David J. Hawes, Mark R. Dadds, *and* Dave S. Pasalich

Abstract

Observational coding involves classifying and quantifying verbal and nonverbal behavioral events or psychological states, irrespective of participants' reports or perceptions. Such coding has been widely used to index the dimensions of diagnostic symptoms associated with various disorders, the contextual dynamics of functional importance to these disorders, and individual differences (e.g., child temperament) and internal processes (e.g., cognitive biases) implicated in pathways to these disorders. We provide an overview of the applications of observational coding strategies in clinical research, and key principles in the design and implementation of observational strategies. Examples are drawn from programs of research that demonstrate the theory-driven use of observation, often in the context of multimethod measurement. We also focus specifically on observational measurement in intervention designs, and make recommendations regarding the role of observation in addressing research challenges associated with emerging models of psychopathology.

Key Words: Observational coding, direct observation, parent–child interaction, relationship dynamics

Introduction

The rich evidence base available to us as clinical practitioners and behavioral scientists owes much to direct observation. Observational coding involves classifying and quantifying verbal and nonverbal (e.g., motor actions, expressed affect) behavioral events or psychological states, irrespective of participants' reports or perceptions (Dishion & Granic, 2004; Heyman & Slep, 2004). In clinical psychology and psychiatry, direct observation has long been an essential diagnostic and therapeutic tool. Although issues related to the use of observation in clinical practice versus clinical science overlap considerably, the primary focus of this chapter is on the latter.

Observational coding is commonly used to measure variables related to psychopathology and dysfunction. First, coding can provide valuable data on the dimensions of diagnostic symptoms associated with disorders. Behavioral avoidance tests, for example, have been used to assess forms of anxious avoidance associated with various anxiety disorders, including specific phobia, agoraphobia, and obsessive-compulsive disorder. In such a test, participants may be exposed to their feared stimuli under controlled settings and observed while they complete as many steps as possible in a graduated series. Ost and colleagues (2001), for example, asked youth with a snake phobia to enter a room where a live snake was enclosed in a glass container and to remove the lid and pick up the snake and hold it for 10 seconds. The percentage of steps the youth accomplished was then observed and recorded.

Second, observation is widely used to measure the contextual variables associated with mental health problems, and in turn to capture the qualities of social contexts that maintain and amplify these problems over time. In such a capacity, observation

allows for the systematic examination of functional relationships between problem behavior and the environment in which it occurs. It is for this reason that observational methods have played an integral role in the early development of behavioral interventions. Indeed, much of what we currently know about risk and protective processes within the relationship contexts of couples, parents and children, peers, and siblings has its origins in data coded from direct observation. As discussed later, observational coding of the features and contextual interactions continues to play a key role in the scientific investigation of problem development, treatment effects, and mechanisms of behavior change.

Third, observational coding is commonly used to assess participant characteristics or individual differences that—while not symptoms of psychopathology per se—may be implicated in pathways to psychopathology. A common example is the use of observation to code dimensions of child temperament, as in Caspi's (2000) longitudinal study of personality continuities from childhood to adulthood. In this study the temperaments of 3-year-olds, coded from direct observation, were found to predict psychological disorders, including depression, antisocial personality disorder, and alcohol dependence, in those individuals at 21 years of age. Driven by emerging models of temperament, lab-based observational paradigms have been developed to index temperament dimensions (e.g., Goldsmith & Rothbart, 1996; Kochanska, Murray, & Harlan, 2000). Such paradigms have featured increasingly in clinical research as interest in developmental models of psychopathology has grown.

Fourth, observational coding has at various times been used as a means to indirectly index covert, internal events—typically cognitive processes—that are not amenable to measurement via direct observation or self-report. Ehrmantrout and colleagues (2011), for example, videotaped depressed adolescents and their parents in problem-solving interactions, and coded the affect expressed by parents. The adolescents also viewed these recordings in a video-mediated recall procedure in which they were required to identify their parents' emotions in 20-second intervals. By analyzing discrepancies between adolescents' subjective ratings of parent affect and the independently coded observations of this affect, researchers were able to identify the emotion recognition deficits that characterized these adolescents.

Direct observation has been described as the most effective method available for obtaining ecologically valid information about behavior (Barkely, 1997) and carries distinct advantages over other strategies in the empirical study of psychopathology and its treatment. These advantages have become increasingly apparent as evidence regarding the nature of psychopathology has emerged. In the past two decades, experimental research has focused increasingly on cognitive aspects of psychopathology, with findings emphasizing various biases in information processing across a range of clinical populations (e.g., Dobson & Kendall, 1993). For example, evidence shows that the patterns of overreactive parenting that are associated with childhood conduct problems are also associated with parental deficits in the encoding and appraisal of child behavior. Such parents appear more likely to notice negative, relative to positive, child behavior, and to view neutral or positive child behavior as problematic (e.g., Lorber, O'Leary, & Kendziora, 2003). At the same time, there is growing evidence that the behaviors of functional importance to many disorders occur in *overlearned* patterns (Dishion & Stormshak, 2007; Gottman, 1998). That is, as a result of frequent repetition they are often performed somewhat automatically, outside of awareness. For example, the reactive actions of distressed couples—often corroding relationship quality through escalating cycles of escape-conditioning—are likely to be enacted with the same automaticity of other overlearned behaviors such as driving a car or reading. The implication of such evidence for the topic at hand is that the cognitive and behavioral processes that underlie various clinical problems have the potential to confound self-reports of those problems and their components. Alternatively, data coded from direct observation are neither limited by participants' explicit awareness of behaviors nor contaminated by the perceptual biases that may color their reports. Such advantages may often offset the significant costs of adopting observational measures in clinical research, and have often been cited by researchers who have (e.g., Sheeber et al., 2007).

The importance of observational strategies to clinical research can also be appreciated in relation to the emphasis on the social environment in many research questions concerning the development and treatment of psychopathology. Causal models of psychopathology have long emphasized environmental risk factors in the prediction of dysfunction, and decades of research have shown that the most proximal of these often operate through social mechanisms. Historically, this research has focused primarily on family dynamics and has produced

models that have been translated into widely disseminated interventions. It is in the field of childhood conduct problems that family dynamics have probably been investigated most extensively (see Hawes & Dadds, 2005a, for a review), and the most significant examples of this research (e.g., Patterson, 1982) have often relied heavily on observational methods. More recently, observational data have informed increasingly sophisticated conceptualizations of family interactions in pathways to child anxiety (e.g., Dadds, Barrett, Rapee, & Ryan, 1996; Dubi et al., 2008; Hudson, Comer, & Kendall, 2008; Suveg, Sood, Barmish, Tiwari, Hudson, & Kendall, 2008).

Research into peer relationships has provided evidence that contexts outside of the family play a significant role in pathways to a range of disorders across childhood and adolescents. In a review of the emerging evidence, Dishion and Tipsord (2011) identified "peer contagion" as a mechanism through which environments shape and amplify not only problems such as aggression and drug use, but also depression and eating disorders. Importantly, progress in this area has demonstrated that the specific dynamics through which peers confer risk are often subtle and not detectable through methods other than the coding of the behavior stream (Dishion & Granic, 2004). It is clear that direct observation provides a unique window on individuals in the context of—and in relation to—their ecologies, and is a strategy that remains essential to answering many of the questions surrounding the roles that environments play in problem trajectories.

Preliminary Issues in the Use of Observation

Accessibility

In fields where dysfunction is associated with overt or "public" behavioral events, observational methods have proliferated. A noteworthy example is the field of childhood externalizing problems. Diagnostic features of presentations such as oppositional defiant disorder (e.g., often loses temper) can be readily operationalized in terms of observable behavior (e.g., crying, kicking), as can the contextual variables that often accompany them (e.g., aversive and ineffective parenting practices such as criticism and vague instructions). Established systems for observational coding in this area are many and have been subject to psychometric research (see Aspland & Gardner, 2003). Likewise, an extensive range of observational methods have been developed to investigate the relationship problems of distressed couples (see Kerig & Baucom, 2004). Conversely, there are numerous clinical problems—ranging from psychotic disorders to common mood and anxiety disorders—that are characterized by largely internal or "private" symptoms (e.g., hallucinations, feelings of hopelessness and fear). Research in such areas has relied far less on observational measurement, and relatively few observational coding systems have been developed specifically to investigate such problems.

Why are observational strategies used to study some disorders more than others? The difference can be understood in terms of *accessibility*—the notion that some events are more amenable to direct observation than others. Accessibility has been defined as the likelihood that an environmental or behavioral event exists, coupled with the difficulty of reliably observing it (Johnston & Pennypacker, 1993). It is self-evident as to why fields of research concerned with those more public events have typically favored observational methods more than those concerned with events and processes that are largely private. However, public behavior may nonetheless be associated with topographic characteristics that present challenges related to accessibility. Some behaviors may occur too infrequently (e.g., seizures) or too briefly (e.g., some motor tics) to be observed reliably. Alternatively, the contexts in which they occur may not be conducive to observation (e.g., stealing). This issue is seen in "setting events"—events that are functionally related to behaviors of interest, yet occur in contexts that are distally removed from those in which the behaviors occurs. For example, incidents of bullying experienced by a child while traveling to school may function as an important setting event for that child's aggression toward peers in the school playground later in the day. Accessibility is a crucial concern and should be considered carefully when deciding upon the use of direct observation as a measurement strategy. However, such a decision should also be informed by an awareness of the strategies by which issues of accessibility may be minimized—as addressed throughout the following sections.

Accessibility is not an inherent characteristic of specific events or behaviors, but rather is determined by multiple variables associated with the likelihood that an observer will be able to detect them. As such, accessibility may be an issue when established observational systems are not available to assess particular constructs, or simply that appropriate training in those systems is not available to coders. Likewise, observational procedures themselves

may compromise accessibility through participant reactivity. For example, family members may react to the physical presence of observers in the family home by inhibiting certain behaviors that may otherwise feature in typical interactions.

Representativeness

For observational data on a specific set of behaviors to inform empirical conclusions, a representative sample of that behavior must be collected. In other words, observational assessment aims to capture the typical behavior of participants in the setting of interest. The challenge can be likened to the process of conducting a clinical assessment interview. It may be easy to get a client talking about the issues he or she is presenting with; however, only by asking the right questions, and in the right way, will an interview elicit the specific information necessary to formulate a reliable diagnosis. Likewise, the potential for observational measurement to capture the typical behavior of an individual will depend on considerations such as how much of that behavior is sampled, on how many occasions it is sampled, and under what conditions the sampling occurs. For example, early observational research into marital problems found that the interactions of distressed couples were indistinguishable from those of nondistressed couples when using standardized analogue tasks (e.g., Birchler, Weiss, & Vincent, 1975). It was only when such couples were observed discussing sensitive issues in their own relationships that features uniquely associated with relationship quality could be detected (e.g., Gottman, 1979).

Settings for Observational Measurement

The conditions under which behavior is observed in clinical research typically span a continuum ranging from unconstrained naturalistic observation to tightly controlled analogue tasks in laboratory or clinic settings (Hartmann & Wood, 1990). Naturalistic observation typically refers to coding behavior outside of the laboratory, in the "real world" (see Dishion & Granic, 2004, for a review). Common locations for such observation include the family home, classrooms, and schoolyards; however, they may in principle be conducted anywhere. Given the presence of either an observer or camera, and the ethical requirement that participants are aware that their behavior is being recorded, consumers of such research have often queried the extent to which naturalistic observation does truly capture real-world behavior. Social desirability is a common concern, considering the range of behaviors that are likely to be of interest to clinical researchers (e.g., harsh parenting practices, couples' hostility). Such issues have also been examined empirically in measurement research (see Gardner, 2000, for a review of parent–child interactions). Experimental studies have been conducted in which the intrusiveness of recording equipment has been manipulated (e.g., Jacob, Tennenbaum, Seilhamer, Bargiel, & Sharon, 1994) or participants have been instructed to intentionally fake particular types of behavior (e.g., Johnson & Bolstard, 1975; Patterson, 1982). Rather than suggesting that participants inhibit socially undesirable behavior during naturalistic observation, such studies have provided impressive support for the reliability of such methods. Data from these studies indicate that family interactions vary little based on the presence or absence of an observer, and that although participants are easily able to fake "bad" (e.g., critical and reactive couples communication), they are generally unable to fake "good" (e.g., mutually supportive couples communication).

What about the impact that location may have on observed participant behavior? This question has been the subject of measurement research, with studies comparing participant behavior in home versus laboratory/clinic settings. Mothers and children have been found to exhibit higher rates of various behaviors in the clinic setting compared to home (Zangwill & Kniskern, 1982). There are also some data to suggest that these respective settings may differ in the extent to which they bias participants toward positive versus negative behavior. Findings regarding the direction of such effects vary somewhat across different populations and paradigms. For example, when observed in the laboratory, mothers have been found to be more active and responsive to their infants and more interactive and helpful and less restrictive in parent–child observations (Jacob, Tennenbaum, & Krahn, 1987; Moustakas, Sigel, & Schalock, 1956) compared to in the home. Married couples have likewise been found to engage in more positive emotional interactions in the laboratory setting compared to the home (Gottman, 1979, 1980), whereas families have been found to exhibit more positive interactions during decision-making tasks conducted in the home setting (O'Rourke, 1963).

Naturalistic Observation

Clinical psychologists have often aimed to observe behavior in its natural context where possible, based on the assumption that *in vivo* observation potentially provides the highest-quality data

(Cone, 1999). Naturalistic observation has often been utilized in studies that test the therapeutic effects of modifying contextual variables on participant outcomes (e.g., Raver et al. 2009), and those concerned with the specific contexts in which behaviors occur. For example, Snyder and colleagues (2008) used a multimethod strategy to assess child antisocial behavior in each of three social ecologies (home, classroom, and school playground) to index its cross-context stability. School playground data were collected by observing the behavior of participating children in this setting on 10 separate occasions between the ages of 5.3 and 6.8 years. On each occasion, the behavior of each child was observed and coded for 5 minutes in relation to peer aggression and covert antisocial behavior.

The major advantage to naturalistic observation is the likelihood that data will generalize to the real world (Mash & Terdal, 1997). However, as noted by Dishion and Granic (2004), much of what occurs in the real world does not provide informative data on the functional dynamics of psychopathology and adjustment. The authors point out that observing discordant couples or families throughout the course of their day is likely to reveal little about the interpersonal process related to their conflict, as the interactions associated with conflict are often avoided (for that reason) by the individuals involved. As such, naturalistic observation often requires researchers to place some restrictions or structure on the behavior of those individuals being observed. The aim of this structure is generally to elicit the most meaningful behaviors. Such restrictions may be somewhat minimal, involving home visits during which a parent and child are asked to engage in unstructured play using age-appropriate toys for a set period of time. Alternatively, the observation may be scheduled around events in a family's daily routine that are "high risk" for behaviors of interest, as is often the case with mealtimes for young oppositional children. In either case, at least minimal restrictions are likely to be imposed by the researcher, such as asking family members to remain in two adjacent rooms, and leaving televisions and telephones turned off (e.g., Maerov, Brummet, & Reid, 1978).

Research laboratories are often the preferred settings for scheduling carefully controlled observations, allowing for access to equipment such as digital recording facilities. However, naturalistic observation may also involve a high degree of structure. An example of this is the observational assessment used by Trentacosta and colleagues (2008) to examine relations among cumulative risk, nurturant and involved parenting, and behavior problems across early childhood. During a home visit, children and caregivers were videotaped in a series of highly structured tasks designed to sample common family scenarios and elicit a range of child and parent behaviors. These included a cleanup task (5 minutes), a delay of gratification task (5 minutes), teaching tasks (3 minutes each), the presentation of two inhibition-inducing toys (2 minutes each), and a meal preparation and lunch task (20 minutes). Another example is the structured family discussion paradigm used by Dishion and Bullock (2001). Once again, during a home visit parents and adolescents were videotaped engaging in discussions on a series of set topics. Not only were the topics of discussion structured (ranging from planning an activity for the following week, to parental monitoring of peer activities, and norms for substance use), but also was the participation of various family members, with siblings included selectively in specific discussions. These structured discussions were coded for parent–adolescent relationship quality, and problem-solving and parenting practices related to the management and monitoring of adolescent behavior. Importantly, increasingly sophisticated and affordable technology for mobile digital recording may lead to increases in the conduct of naturalistic observations in clinical research in the coming years.

Analogue Observation

In contrast to the real-world contexts of naturalistic observation, analogue observations are conducted in artificial settings that are often far removed from those in which behaviors of interest are typically performed. Behavioral laboratories and psychology clinics are often the venues of choice for observation of this kind, which typically involves conditions that are designed, manipulated, or constrained by researchers (see Heyman & Slep, 2004). Although issues of cost and convenience often underlie the adoption of analogue observations over naturalistic ones, analogue methods of observation present a number of distinct advantages. First and foremost, analogue observation provides a high degree of control over the conditions under which behavior is observed, allowing researchers to standardize such conditions across participants. Like the structured observational tasks conducted in naturalistic settings, analogue observations are typically structured in order to elicit specific behaviors of interest. Again, these may be low-frequency behaviors or those that

are difficult to view in their natural context for other reasons. Importantly, however, the types of restrictions that can be placed on participant behavior in the laboratory are potentially more complex than those that are often possible in naturalistic setting. These controlled conditions may also be manipulated in experimental designs, allowing researchers to test predictions concerning the effects of specific stimuli or events on participant behavior.

Such an approach was used by Hudson, Doyle, and Gar (2009) to examine child and parent influences on dimensions of parenting associated with risk for child anxiety. Mothers of children with anxiety disorders and mothers of nonclinical children were videotaped interacting during a speech-preparation task with a child from the same diagnostic group as their child (i.e., anxious or nonanxious) and with a child from the alternative diagnostic group. Maternal behavior was then coded in terms of overinvolvement and negativity. It was found that when interacting with children other than their own, mothers were observed to be more involved with anxious children compared to nonclinical children. The use of analogue observation in this design allowed the researchers to identify potentially important bidirectional effects between child anxiety and parenting practices.

The importance of observational context has been emphasized in a number of studies. Dadds and Sanders (1992) compared observational data collected through home-based free parent–child interactions versus clinic-based structured mother–child problem-solving discussions, in samples of depressed and conduct-disordered children. Parent–child behaviors observed during unconstrained home interactions showed relatively little convergence with behavior observed in the clinic-based problem-solving tasks. No relationship was seen between children's behavior across each of the respective settings. Maternal behavior was somewhat more consistent, with mothers' depressed affect during clinic-based problem solving related to aversive behavior in the home for mothers of depressed children. Likewise, angry affect during problem solving was related to aversive behavior in the home for mothers of conduct-disordered and comparison children. In terms of predictive validity, observations of depressed children and their mothers in home-based interactions correctly predicted child diagnoses in 60 percent of cases. This was compared to only 25 percent accuracy based on behavior observed during clinic-based problem solving. Conversely, accuracy of classification for conduct-disordered children based on observations of problem solving and home-based free interaction was 72 percent for each. The design of this study precluded the authors from disentangling effects related to setting (home vs. clinic) from those related to the restrictions imposed on participants (relatively unrestricted naturalistic vs. structured analogue). However, such findings nonetheless demonstrate the potential for these respective observational contexts to provide unique diagnostic information related to different disorders. These findings also suggest that the inclusion of multiple observational contexts may produce the most comprehensive behavior data related to clinical risk.

Support for these assumptions can be found in subsequent studies, such as the recent investigation of pathways to adolescent depression reported by Allen and colleagues (2006). Using two separate analogue observations, adolescents' behavior was coded on dimensions of autonomy and relatedness, first in the context of the mother–child relationship and then in the context of peer relationships. In the first observation adolescents and their mothers participated in a revealed-differences task in which they discussed a family issue (e.g., money, grades) that they had separately identified as an area of disagreement. In the second observation the adolescent was videotaped interacting with a close friend while problem solving a fictional dilemma. Using a longitudinal design, the authors showed that adolescents' behavior with their mothers (associated with undermining autonomy and relatedness) and peers (associated with withdrawn-angry-dependent interactions) both explained unique variance in growth of depressive symptoms. The prediction of problem trajectories would therefore have been reduced had the behavior been observed in the context of only one of these relationships.

Interesting developments concerning the importance of context to observational measurement have come from studies adopting dynamic systems (DS) theory—a mathematical language used to describe the internal feedback processes of a system in adapting to new conditions. Granic and Lamey (2002) used a problem-solving paradigm to observe parent–child interactions in a sample of boys with clinic-referred conduct problems, with and without comorbid anxiety/depression. Guided by the assumptions of DS theory, the observational procedure incorporated a perturbation—an event intended to increase pressure on the parent–child dyad and trigger a reorganization of their behavioral system. A core premise of DS approaches is that perturbations expose the characteristics of a system as it moves away from and

back to a stable equilibrium. The specific perturbation employed was a knock on the laboratory door to signal that the allotted time for the problem-solving discussion was almost over, and that a resolution was needed. The rationale for this perturbation included the DS premise that only by perturbing a system can the full range of behavioral possibilities therein be identified. The authors found that parent–child interactions in the two groups differed only after the perturbation. Specifically, during the initial period of the problem-solving discussion, parents in both groups exhibited a permissive style in responding to aversive behavior in their children. However, following the perturbation, only those dyads involving children with comorbid internalizing problems shifted to a style of interaction characterized by mutual and escalating criticism and hostility. The finding that these parent–child dyads were more sensitive to the effects of the perturbation than those of pure externalizing children was interpreted as evidence of structural differences between these respective types of dyads at a systemic level. Importantly, these findings also demonstrate that important classes of behavior may at times be observable only by placing pressure on participants through the systematic manipulation of contextual cues.

Approaches to Coding Behavior

Observational coding strategies vary considerably in terms of the specificity or precision with which behavioral codes are operationalized. Molecular (or microsocial) coding systems are the most intensive, specifying discrete, fine-grained, behavioral units (e.g., eye contact, criticize, whine). At the other end of the spectrum are molar (or global/macrolevel) coding systems, based on more inclusive behavioral categories. For example, the code "whine" is much more specific (or molecular) than the more global code "oppositional behavior." Researchers interested in short-term patterns of behavior per se—such as those testing predictions from operant theory—have typically favored the coding behavior at the molecular level. Such coding is likely to be of particular value when these patterns of behavior are associated with potentially important variations across time and contexts, and can be accounted for by social influence. Conversely, when behavior is used merely as a "sign" of an underlying disposition or trait, or researchers are interested in events and extended processes that occur over longer time scales, the coding of more molar or global categories may be of greater value (Cone, 1999; Dishion & Granic, 2004).

Microsocial Coding Systems

Observational systems concerned with describing behavior at a molecular level typically do so by coding discrete behaviors into mutually exclusive categories. These categories are operationalized in concrete behavioral terms that allow them to be coded with minimal inference. The term "microsocial" has traditionally been applied to observational coding that is concerned with the order and pattern of behaviors in a stream of observed social interaction (Dishion & Snyder, 2004). The strength of such coding is its potential to describe behavior as it unfolds over time. By representing the moment-by-moment interactions between individuals in the contexts of relationships (e.g., parent–child, peer, spousal, sibling), microsocial coding can capture the relationship processes that underlie dysfunction and adjustment at both individual and systemic levels.

Mircosocial coding was integral to the classic observational studies conducted by Patterson and colleagues at the Oregon Social Learning Center, beginning in the 1960s and 1970s (see Patterson, Reid, & Dishion, 1992). These influential studies examined the moment-to-moment interactions within families of aggressive and oppositional children, and the functional dynamics between these interactions and children's antisocial behavior. Seminal coding systems were developed at this time to capture the microsocial interactions of families (e.g., the Family Process Code; Dishion et al., 1983) and administered live in naturalistic settings—most often the family home. Observational data from this research indicated that three moment-to-moment variables most robustly differentiated the interactions of families of clinic-referred conduct problem children from those of well-adjusted children. The first was "start-up"—the likelihood that a family member would initiate conflict when others were behaving in a neutral or positive manner. The second was "counterattack"—the likelihood that a family member would react immediately and aversively to an aversive behavior directed at him or her by another family member. The third was "continuance"—the likelihood that a family member would continue to act in an aversive manner following the first aversive initiation. Importantly, the moment-to-moment account of these family interactions provided by microsocial observation allowed Patterson (1982) and colleagues to apply social learning theory to family process. Their subsequent conceptualization of "reinforcement traps"—based on escape-avoidance conditioning—forms the basis for the most established interventions

currently available in this area (see Eyberg, Nelson, & Boggs, 2008).

Global Coding Systems

Global or molar coding systems assign codes based on summary ratings of behavior, and often across longer time scales. Codes tend to be few, representing behavioral classes (e.g., negativity, supportiveness, conflict/hostility). The speed and simplicity with which such systems can often be administered appeal to many of the researchers who adopt them. Furthermore, global ratings have often been reported to be correlated highly with microsocial data in studies comprising both. Hops, Davis, and Longoria (1995) found this to be the case for not only the global ratings of parent–child interactions made by trained observers, but also the global ratings made by parents themselves. Heyman and Slep (2004) suggested that global ratings may often be very appropriate in the collection of observational data, given that molecular systems comprising 30+ codes are often collapsed down into composite variables comprising positive, negative, and neutral dimensions for the purpose of statistical analysis. It is important to remember, however, that while the option to create such composite variables is available when raw observational data are captured by molecular codes, global ratings of such dimensions can never be disaggregated into the discrete behaviors that they summarize.

As global or molar systems are often better able than microsocial coding to take the broader context of behavior into account, such ratings have the potential to capture some constructs more appropriately than molecular codes. For example, marital interactions have been coded using global ratings of emotional intensity, conflict tactics, and degree of conflict resolution to investigate the relative effects of parental mood and conflict on child adjustment (Du Rocher Schudlich & Cummings, 2007) and therapeutic change following couples' intervention (Merrileesa, Goeke-Moreyb, & Cummings, 2008). Aksan, Kochanska, and Ortmann's (2006) system for coding mutually responsive orientation (MRO) in the parent–child relationship is another such example. This system was developed to measure attachment-related dynamics in parent–child interactions, based on the aim of characterizing such dynamics using both parent and child data. MRO is coded using global ratings that emphasize the joint aspect of parent–child interaction (e.g., "Interaction flows smoothly, is harmonious; communication flows effortlessly and has a connected back-and-forth quality") over the respective behaviors of either member of the dyad. Such ratings are formulated after observing parent–child dyads in a range of structured contexts, each approximately 10 minutes in duration. The authors contrast this method with traditional attachment paradigms in which both parent and child are typically involved yet only the behavior of the child is coded (Aksan et al., 2006).

One of the main limitations of global ratings is that they do not retain the sequential relations of events, and therefore provide less potential information on the functional dynamics of behavior than microsocial coding. However, there is also evidence that global ratings can capture unique information of functional importance. In particular, such ratings may provide unique information about events that unfold over longer time scales, and capture important outcomes to extended processes (Dishion & Granic, 2004). Driver and Gottman (2004), for example, examined the interactions of newlywed couples over the course of a day. Through the global coding of bids for intimacy and conflict discussions, the researchers were able to draw conclusions about the role of affection in couples' conflict resolution, within the broader dynamics of daily interactions. Based on the potentially unique insights provided by microsocial and global coding approaches, Dishion and Snyder (2004) advised that such methods may complement each other in the same programs of research.

Units of Measurement

To quantify the degree to which an observed behavior is performed, a unit of measurement must be assigned to some aspect of that performance (Johnston & Pennypacker, 1993). In clinical research a number of parameters—or dimensions—of behavior are often indexed for this purpose. The most common of these are frequency, intensity, permanent products, and temporal properties. The aims of research are most likely to be met when the theory-driven conceptualization of variables determines the precise dimensions of behavior through which they are indexed by observation. At the same time, decisions related to such units of measurement must also take into account the topographic characteristics of the behavior of interest. Importantly, the dimensions through which behavior is indexed have implications for other aspects of the coding strategy, as different methods of recording observational data are better suited to capturing different dimensions of behavior.

The *frequency* with which a behavior occurs has traditionally been seen to reflect the strength of that behavior, and is often the simplest dimension to observe for discrete behavioral events—those with an identifiable beginning and end. Frequency can also be one of the simplest to interpret, providing rate indices that can be standardized across various periods of time (e.g., rate per minute, rate per hour). The *intensity* of a behavior refers to the amplitude, force, or effort with which it is expressed. However, as intensity is related largely to the impact that a behavior has on the environment rather than the characteristics of the behavior itself, it can be more difficult to observe than frequency or temporal dimensions of behavior (Kazdin, 2001).

Permanent products are the tangible byproducts or "trace evidence" of behavior (e.g., number of wet bedsheets, number of windows broken, number of chores completed). Although not a measure of behavior per se, these are measures of the result or effect of behavior, and may be of value when a behavior itself is not readily observable but leaves some lasting product that can be obtained or observed (Bellack & Hersen, 1998). It is likely, however, that such data—which can be recorded simply by noting such occurrences—will often be more informative to clinicians than researchers. Unlike the other dimensions of behavior addressed here, it does not provide any information about the form or function of the behaviors that produce such products.

The *temporal dynamics* of behavior may also be related to clinically important processes and can be characterized in various ways. These include duration (the amount of time that elapses while a behavior is occurring), latency (the amount of time that elapses between the presentation of a stimulus and the onset of a response), and interresponse time (the amount of time that elapses between the offset of one response and the onset of another). Piehler and Dishion (2007), for example, observed the interactions of adolescents in a discussion task with a close friend. The simple index of duration of deviant talk bouts was found to differentiate youth with early-onset antisocial behavior, late-onset antisocial behavior, and normative behavioral development.

The temporal dynamics of behavior have proven to be of particular value in the fine-grained analysis of relationship processes, as we shall soon discuss.

Approaches to the Recording of Observational Codes

A range of strategies can be used to record the occurrence, sequence, intensity, and duration of behaviors, and in turn quantify the behavior stream. The strategies we will focus on here are event records, interval-based time sampling methods (partial interval, whole interval, momentary), and those that represent the temporal dynamics of behavior. These strategies present unique pros and cons when applied to different dimensions of behavior. The appropriateness of a specific recording method may also depend on practical considerations related to the setting in which the behavior is recorded. Attempts to capture fine-grained data on behavioral interactions will be of little use if the complexity of the observation strategy prohibits the reliable sampling of that behavior. This may be a particular concern for those increasingly rare studies that rely on the live observation of participants. In such studies, the complexity of the recording strategy will in part determine the demands that are placed on the observer's attention, or observer load. Where coding is completed from digital video/audio recordings, it is often possible to minimize the demands associated with such complexity through repeated viewings of footage, and in some cases the use of commercially available software designed for such purposes. However, coding from video footage may nonetheless become prohibitive if the time spent reviewing such recordings far exceeds the real time that it represents.

Event records involve the recording of each and every occurrence of a behavior in a given period of time. This recording strategy is most useful for discrete behaviors that are low in frequency (e.g., swearing, out of seat in classroom, throwing a tantrum). Such events may be those performed by an individual or a group (e.g., children in a classroom). Event records are relatively simple to design and implement, and carry the advantage of providing relatively complete coverage of the behavior of interest. Event-based data may be converted into rate indices by dividing the observed frequency of behaviors by the amount of time observed, thereby allowing for comparison across variable periods of time. However, for many studies it is not feasible to collect such comprehensive records for the large volumes of behavior that are of potential interest; nor is it necessary in order to examine the dynamics of fine-grained behavior.

Breaking a period of observation into smaller segments or intervals is a common practice that allows large volumes of behavior to be recorded and facilitates the formal evaluation of reliability by allowing for point-by-point comparison between observers. For example, a 30-minute observation period

may be divided into 30 1-minute intervals or 180 10-second intervals. Interval-based recording is most efficient when the length of the observation interval is related to the frequency of the behavior, with high-frequency behaviors recorded using shorter observation intervals and low-frequency behaviors longer intervals. Dishion and Granic (2004) advise that intervals in the region of 10 to 15 seconds typically retain the microsocial quality of behavioral interactions being observed. However, somewhat longer intervals are often adopted due to practical considerations, and various time sampling methods make use of these intervals in different ways. Such intervals serve the purpose of allowing observers to simply record whether or not a behavioral response has occurred, as opposed to recording every instance of that behavior. The raw data recorded in interval-based methods are typically converted into percentages of intervals observed. As such, they represent an estimate of behavior frequencies rather than the absolute frequencies of those behaviors.

Partial interval time sampling is used to record the presence or absence of a behavior if it occurs once or more at any time during a given interval. For example, if a behavior occurs twice in one interval and ten times in the next, both intervals will register the same unit of behavior (behavior present). This common method is useful for high-frequency, brief behaviors that do not have a clear beginning or end. For example, an observer coding a 20-minute parent–child interaction task may record "yes" to each consecutive, 15-second interval in which any designated parent or child behaviors occur. Once recorded, such data can be expressed in terms of the percentage of intervals observed in which any instance of the behavior occurred. Partial interval time sampling is well suited to observations concerned with the frequency of behavioral responses and has been widely used for this purpose. There is some evidence, however, to suggest that this strategy tends to overestimate behavior frequency (Green & Alverson, 1978; Powell et al., 1977). Alternatively, the risk that partial interval recording may underestimate the frequency of high-rate behaviors increases with increases in the length of the recording interval. Partial interval systems may also be used to estimate response duration; however, this is less common and relies on interval length being brief relative to the mean duration of the behavior of interest in order to minimize overestimation (Hartmann & Wood, 1990).

Whole interval time sampling registers behavioral responses that occur throughout the entire length of a given interval. This strategy has traditionally been used most often in research concerned with the duration of behavior, but it can also be used to record estimates of frequency. Whole interval time sampling is most suited to the observation of behaviors that are of lengthy duration or may not have a clearly identifiable beginning or end. Dion and colleagues (2011), for example, used this method to collect observational measures of children's classroom attention in a randomized clinical trial aiming to improve attention and prevent reading difficulties. Participating children were observed in the classroom for 12 consecutive 5-second intervals, and for an interval to be coded as "optimally attentive," the child was required to be correctly seated and oriented toward the relevant teaching stimuli for its full duration. In terms of measurement error, there is some evidence that whole interval time sampling tends to underestimate both the absolute duration and frequency of behavior (Powell et al., 1977).

Momentary time sampling registers the occurrence of a behavior if it is occurring at the moment a given interval begins or ends. It is typically used to provide frequency and duration data. This strategy is suited to long-duration or high-frequency behaviors (e.g., rocking in an autistic child, on-task behavior in the classroom). For example, Brown anc colleagues (2009) used momentary time sampling to code multiple dimensions of children's physical activity in school settings, including level of physical activity (e.g., stationary with limb or trunk movement, vigorous activity), and primary topography (e.g., running, sitting, standing). Observers noted the child's behavior every 30 seconds throughout a 30-minute observation period and assigned such codes based on the behavior occurring at that moment. There is some evidence to suggest that momentary interval sampling tends to underestimate the frequency of behavior, particularly for behaviors that are of short duration (Powell et al., 1977). Gardenier, MacDonald, and Green (2004) compared momentary time sampling with partial interval sampling methods for estimating continuous duration of stereotypy among children with pervasive developmental disorders. While partial interval sampling consistently overestimated the duration of stereotypy, momentary sampling at times both overestimated and underestimated duration. Momentary sampling was found to produce more accurate estimates of absolute duration across low, moderate, and high levels of this behavior.

In contrast to the discontinuous account of behavior recorded by time interval methods,

behavior may also be recorded second by second, in a continuous stream. Interest in the "real-time" temporal properties of social dynamics has grown considerably in recent decades, supported by emerging methods and frameworks for recording and analyzing temporally laden interaction patterns. Recent innovations have allowed researchers to investigate dimensions of social interaction that are inaccessible to methods of behavioral observation that do not capture the temporal quality of real-time change in behavior as it responds to varying environmental demands. Some of the most important developments in this area have focused on the nonlinear dynamics of relationship patterns—often associated with sudden shifts—that are difficult to model using traditional analytic methods.

Such developments include Gottman's (1991) framework for conceptualizing the nonlinear dynamics of close relationships, and the development of methods for the mathematical modeling of relationships based on DS theory (Ryan et al., 2000). Gottman's approach uses a time series of coded observational data to create parameters for each member of a dyad. These parameters are used to identify key patterns of dyadic interaction, and the trajectories toward these patterns can be analyzed to reveal the underlying dynamics of the system. Gottman and colleagues used this approach to model the dynamics of marital communication, showing that these dynamics can predict those couples who divorce and those who will remain married (Gottman, Coan, Carrere, & Swanson, 1998). This method has also been applied to children's interactions in peer dyads. For example, Gottman, Guralnick, Wilson, Swanson, and Murray (1997) modeled the observed peer interactions of children with developmental delays in this way to examine the role that peer ecology plays in the emotion regulation of such children.

DS principles have also formed the basis for other innovations in the recording and analysis of observed relationship interactions. A particularly noteworthy example is the state space grid (SSG). Developed by Lewis, Lamey, and Douglas (1999), the SSG is a graphical and quantitative tool that can be used to create a topographic map of the behavioral repertoire of a system (e.g., parent–child dyad). It works by plotting the trajectory (i.e., sequence of emotional/behavioral states) on a grid similar to a scatterplot, which is divided into a number of cells. The coded behavior of one member of the dyad (e.g., the child) is plotted on the *x* axis and the other member's (e.g., parent) on the *y* axis. Each point (or cell) on the grid therefore represents a simultaneously coded parent–child event, or "stable state" of the dyad. Any time a behavior changes, a new point is plotted and a line is drawn connecting it to the previous point. Thus, the grid represents a series that moves from one dyadic state to another over the course of an interaction. Figure 8.1 shows a hypothetical trajectory representing 10 seconds of coded parent–child behavior on a SSG (Hollenstein et al., 2004). This behavioral sequence begins with 2 seconds in negative engagement/negative engagement

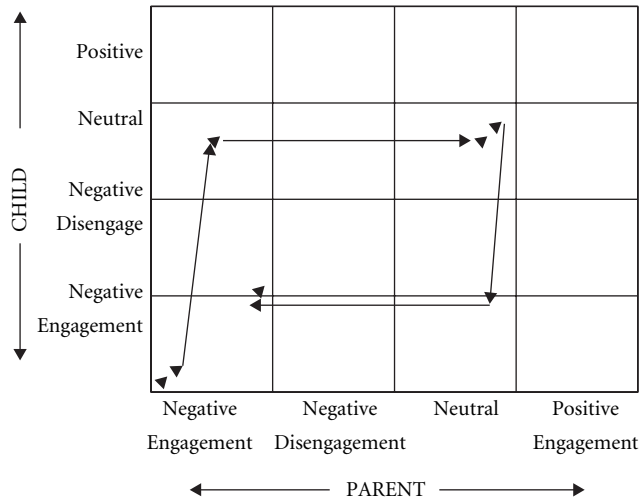


Figure 8.1 Example of a state space grid with a hypothetical trajectory representing 10 seconds of coded behavior, one arrowhead per second. Plotting begins in the lower left part of the cell and moves in a diagonal as each second is plotted, ending in the upper right (Reprinted with permission from Hollenstein et al., 2007).

and is followed by 2 seconds in negative engagement/neutral, 3 seconds in neutral/neutral, 1 second in neutral/negative engagement, and 2 seconds in negative engagement/negative engagement.

This method can be used to record and examine several coexisting interaction patterns and explore movement from one to the other in real time. In DS terms, the moment-to-moment interactions of the dyad are conceptualized as a trajectory that may be pulled toward certain attractors (recurrent behavioral patterns or habits) and freed from others. Using SSG, attractors are identified in terms of cells to which behavior is drawn repeatedly, in which it rests over extended periods of time, or to which it returns quickly. This temporally sensitive method can be used to examine whether behavior changes in few of many states (i.e., cells) or regions (i.e., a subset of cells) of the state space. It is also possible to track how long a trajectory remains in some cells but not others, and how quickly it returns or stabilizes in particular cells (Dishion & Granic, 2004; Granic & Lamey, 2002).

Novel studies using SSGs have contributed significantly to the clinical literature in recent years. A major focus of such studies has been on the structure or the relative flexibility versus rigidity that characterizes the exchanges within dyadic relationships (e.g., parent–child, husband–wife, adolescent–peer). For example, Hollenstein, Granic, Stoolmiller, and Snyder (2004) applied SSG analysis to code observations of parent–child interactions in the families of kindergarten children to examine whether individual differences in dyadic rigidity were associated with longitudinal risk for externalizing and internalizing problems. Parent–child dyads were observed in 2 hours of structured play and discussion tasks and common components of family routines (e.g., working on age-appropriate numeracy and literacy, snack time). Videotaped observations were coded using the Specific Affect (SPAFF) coding system (Gottman, McCoy, Coan, & Collier, 1996). In this system codes are based on a combination of facial expression, gestures, posture, voice tone and volume, speech rate, and verbal/motor response content to capture integrated impressions of the affective tone of behavior. SSG data indicated that high levels of rigidity in parent–child interactions were associated primarily with risk for externalizing problems, predicting growth in such problems over time. The parent–child dyads of well-adjusted children were found to flexibly adapt to context change and display frequent change in affect, whereas dyads of children at risk for externalizing problems exhibited fewer affective states, a greater tendency to remain in each state, and fewer transitions among these states (Hollenstein et al., 2004).

In research examining the role of peer dynamics in the development of antisocial behavior, SSGs have been applied to observed peer interactions to derive measures of dyadic organization or predictability. Dishion, Nelson, Winter, and Bullock (2004) investigated the organization of peer interactions among antisocial and non-antisocial boys. The predictability of dyadic interactions was indexed by calculating logged conditional probabilities of verbal behavior between members of the respective dyads. Findings indicated that adolescent boys who engaged in the most highly organized patterns of deviant talk (i.e., breaking rules and norms) were the most likely to continue antisocial behavior into adulthood.

The potential flexibility with which SSGs can be applied in various research designs is a clear strength of the method, allowing researchers to derive continuous time series as well as categorical and ordinal data for analysis. Furthermore, unlike sequential analysis, this technique does not rely on base rates of behavior to identify important interactional patterns (Dishion & Granic, 2004). It is also evident that DS approaches have the potential to inform observational research based on numerous developmental and clinical theories. For example, given the capacity for DS methods to capture the structure of dyads, it has been suggested that they may be suited to the investigation of attachment dynamics. Specifically, SSGs could potentially represent secure family dynamics in terms of flexible, nonreactive, and synchronous interactive patterns that are "organized" as coordinated and mutual action–reaction patterns (Dishion & Snyder, 2004).

Observational Measurement in Intervention Research

Direct observation has long been a cornerstone of behavioral therapy and was associated with major developments in intervention science across the second half of the twentieth century. Early landmark stimulus control studies relied heavily on observation for the purpose of behavior analysis. Referrals such as aggressive children were observed in family interactions to identify antecedents and consequences of target behaviors, which researchers systematically modified to observe the effects this produced on behavioral responses (e.g., Patterson, 1974). Observational methods have since played a significant role in the development of the many evidence-based treatments that have grown out of

this tradition, and have informed empirical research concerned with numerous aspects of intervention.

Intervention research is generally concerned with questions related to both the efficacy of treatments in producing clinically meaningful change and the mechanisms through which this change is produced. Snyder and colleagues (2006) identified five core elements that need to be defined and measured in intervention trials to clearly answer such questions. Importantly, each of these core elements presents distinct implications for observational measurement. The first core element in this model is the transfer of skills from a training specialist to a training agent. Training agents may include parents, as in the case of parent training programs, or teachers, as in the case of school-based interventions. The key issues in this element concern the agent's acquisition of the skills that are needed to deliver the intervention to the client participant, who would be the respective children in both of these examples. Snyder and colleagues (2006) suggested that for interventions in which such training involves clearly specified sets of therapist behaviors (e.g., queries, supportive statements, instructions, modeling, role playing, feedback, etc.), observation is often advantageous over other methods in the measurement of this therapeutic process. Patterson and Chamberlain (1994), for example, sequentially coded the behaviors of therapists (e.g., "confront," "reframe," "teach") and parents (e.g., "defend," "blame") observed during a parent training intervention and used these data to conduct a functional analysis of client resistance.

The second core element identified by Snyder and colleagues (2006) concerns the quality of the intervention agent's implementation of the treatment with the client participant. In the many trials that have evaluated parent training interventions for conduct problems in young children, formal observations of parent–child interactions before and after skills training have been common. Numerous observational systems have been developed in association with specific parent training programs, allowing researchers to code parents' implementation of these skills as prescribed by specific programs. In Parent–Child Interaction Therapy (McNeil & Hembree-Kigin, 2010) this is achieved using the Dyadic Parent–Child Interaction Coding System (DPICS; Eyberg, Nelson, Duke, & Boggs, 2004)—an extensive coding system that was developed largely for this purpose. This means that the same behaviors may be classified as negative in one coding system and neutral/positive in another, depending on the content of such programs. For example, in the DPICS a directive is coded as a Command only if it is worded in such a way that it tells a child what *to* do; directives that tell the child what *not to* do are coded as Negative Talk—one of the main codes operationalizing aversive/ineffective parenting behaviors. In contrast, other widely used systems for coding parent–child interactions based on closely related models retain the neutral Command code for both kinds of directives (e.g., the Family Process Code; Dishion et al., 1983).

In our own research we have observed parents' implementation of parent training skills for various purposes, including the investigation of child characteristics predicted to moderate the effects of parenting intervention on child outcomes. Hawes and Dadds (2005b) examined the association between childhood callous-unemotional (CU) traits (i.e., low levels of guilt and empathy) and treatment outcomes in young boys with clinic-referred oppositional defiant disorder whose parents participated in a parent training intervention. Parents' implementation of the specific skills taught in this program, including positive reinforcement of desirable behavior, use of clear concrete commands, and contingent, nonreactive limit setting (Dadds & Hawes, 2006), was coded live in the family home using a structured play task and a dinner observation. In doing so, we were able to show that CU traits uniquely predicted poor diagnostic outcomes at 6-month follow-up, independently of any differences in implementation of the intervention by parents of children with high versus low levels of CU traits.

The third core element in Snyder and colleagues' (2006) model concerns change in client behavior across the intervention sessions. Examining this change in relation to change in the actions of the training agent can provide evidence of the mechanisms through which the intervention is producing behavior change. As noted by Snyder and colleagues (2006), the value of using observation in measuring this element may be reduced when client change is associated largely with covert processes (e.g., the formation of explicit intensions). However, in many forms of psychosocial intervention, such change is accessible to observation. For example, Hawes and Dadds (2006) used observation as part of a multimethod measurement strategy to examine change in this way in the context of a parent training trial for child conduct problems. Child behavior was coded from naturalistic observations conducted at the commencement and conclusion of the intervention. A self-report measure of the parenting practices

targeted in the intervention was also completed by parents at the same assessment points. These self-report data were used to assess dimensions of parenting such as inconsistent discipline and parental involvement, which can be difficult to capture in brief periods of observation. Change in these self-reported parenting domains was significantly associated with change in observations of oppositional behavior across the intervention, consistent with the theoretical mechanisms of the clinical model (Dadds & Hawes, 2006).

Research into mechanisms of change has traditionally relied on treatment trials in controlled settings, but researchers have begun to focus increasingly on such processes in real-world (community-based) settings. Gardner Hutchings, Bywater, and Whitaker (2010) recently examined mediators of treatment outcome in a community-based trial of parent training for conduct problems, delivered to the families of socially disadvantaged preschoolers. Like Hawes and Dadds (2006), Gardner and colleagues (2010) used observation as part of a multimethod measurement strategy to overcome the problem of shared method variance. Parenting practices were measured through direct observation of parent–child interactions in the family home, with the DPICS used to code frequencies of parenting behaviors that were then collapsed into summary positive (e.g., physical positive, praise) and negative (e.g., negative commands, critical statements) variables for analysis. Mediator analyses showed that the effects of the intervention on change in child problem behavior was mediated primarily by improvement in positive parenting rather than reductions in harsh or negative parenting (Gardner et al., 2010).

Granic, O'Hara, Pepler, and Lewis (2007) examined an intervention for externalizing problems in a community-based setting, using observation to investigate mechanisms of change related to different parent–child interactions. The authors used the SSG method to examine the changes in parent–child emotional behavior patterns that characterized children who responded positively to the intervention versus those who failed to respond. Using a problem-solving analogue observation conducted pretreatment and posttreatment, SSGs were constructed to quantify previously unmeasured processes of change related to the flexibility of the parent–child dyad. The children showing the greatest response to treatment were those whose families exhibited the greatest increases in flexibility—as indexed by SSGs showing increases in the number of times they changed emotional states, the breadth of their behavioral repertoire, and decreases in the amount of time they spent "stuck" in any one emotional state. Conversely, the interactions of nonresponders became more rigid across treatment. The authors concluded that rigidity is amenable to change through family-based cognitive-behavioral intervention, and that change in flexibility may be one mechanism through which improvements in problem behavior are produced.

The fourth core element identified by Snyder and colleagues (2006) relates to short-term or proximal (e.g., posttreatment) outcomes of treatment. The possibility that participants in intervention research will report symptom reductions simply as a function of receiving an intervention is widely recognized and is generally addressed where possible using a randomized controlled trial design (see Chapter 4 in this volume). Reporter biases and method effects both have the potential to confound the measurement of treatment effects, and observational measurement has been shown to be a highly effective means of minimizing such error. For example, in a study by Dishion and Andrews (1995), parents were found to report large reductions in their adolescents' problem behavior regardless of their random assignment to active intervention versus control conditions; analyses of the coded observations of parent–adolescent interactions, however, revealed that reductions in conflict behavior were specific to conditions that actively promoted behavior change.

Additionally, the measurement of some treatment outcome variables may be achieved more sensitively through direct observation than other forms of measurement. For example, Stoolmiller, Eddy, and Reid (2000) collected multimethod data to evaluate the effects of a school-based intervention for physical aggression in a randomized controlled trial design. Of all the extensive self-report measures collected, only the coded observations of children's playground behavior were sensitive to the effects of the intervention. In another similar example, Raver and colleagues (2009) used observations and teacher ratings of preschoolers' classroom behavior to evaluate the effects of a classroom-based intervention to reduce behavior problems in children from socioeconomically disadvantaged (Head Start) families. Children's externalizing (disruptive) and internalizing (disconnected) behaviors were observed by coders in 20-minute blocks during the course of a school day. Socioeconomic risk was found to moderate the effects of the intervention on child outcomes, but only in the analyses using the observational data. Such findings reinforce the

value of including observation as part of a multimethod measurement strategy in intervention research (Flay et al., 2005).

In the example of parent training interventions for child conduct problems, short-term outcomes (e.g., reduced oppositional defiant disorder symptoms) have been shown to then contribute to reduced risk for delinquency, drug use, and school failure (Patterson, Forgatch, & DeGarmo, 2010). It is this distal change, represented by long-term outcomes such as these, that is the focus of the fifth and final core element in Snyder and colleagues' (2006) model. Such outcomes are seen to reflect more global and enduring reductions in dysfunction, or increases in capacities and resilience. In contrast to the earlier elements in the model, Snyder and colleagues (2006) suggest that observational methods are often not appropriate to index such outcomes, recommending instead that approaches such as multi-informant self-report measures may provide superior data. As reviewed here, observational coding can serve multiple purposes in intervention and prevention designs. Importantly, observational data are most likely to inform innovations in intervention science when collected in the context of a theory-driven multimethod measurement strategy.

Reliability and Validity

The process of establishing adequate reliability in observational coding is one of the first essential steps in research with such methods, whether this involves the design and development of a novel observational strategy or the implementation of an established coding system. This typically requires that a team of observers are trained until conventional criteria for interobserver agreement are reached. The process for such training often involves intensive practice, feedback, and discussions focused on example recordings, and it may take days to months depending on the complexity of the coding system. Formal calculations of reliability are derived from the completed coding of a sample of recordings by multiple observers who are unaware of each other's results. These calculations may range from a simple index of agreement such as percent agreement for occurrence and nonoccurrence of observed events, through to intraclass correlation coefficients, and an index of nominal agreement that corrects for chance agreement (i.e., kappa). Ongoing training beyond such a point is also advisable to reduce observer drift over time. According to Bakeman and Gottman (1997), focusing on interobserver agreement serves three critical goals. First, it ensures that observers are coding events according to the definitions formulated in a coding manual; second, it provides observers feedback so as to improve their performance; and third, it assures others in the scientific community that observers are producing replicable data.

Considerably less emphasis is typically placed on issues of validity in observational research. This follows from the notion that observational measurement of behavior does not involve the measurement and interpretation of hypothetical constructs. As such, observational data has been viewed as axiomatically valued and its validity taken at face value (Johnston & Pennypacker, 1993). The validity of observational data has nonetheless been examined in measurement research, with evidence of various forms of validity available from a range of multimethod studies (see Cone, 1999). Hawes and Dadds (2006), for example, examined associations between observational measures of parent behaviors coded live in the home setting and parent self-reports of their typical parenting practices on the Alabama Parenting Questionnaire (APQ; Shelton et al., 1996). Evidence of convergent validity was found, with moderate correlations seen between conceptually related observational and self-report variables. For example, observed rates of aversive parenting correlated positively with parent-reported use of corporal punishment, and likewise, observed rates of praise with parent-reported use of positive parenting practices (Hawes & Dadds, 2006).

Future Directions

Before considering future directions for observational coding, it is worth reflecting on trends in the popularity of such strategies. Namely, it seems that the observation of behavior—once ubiquitous in clinical research—has begun to disappear in recent decades (Dishion & Granic, 2004). This trend is not unique to clinical psychology, with a marked decline in the use of observational measurement also noted in other fields such as social psychology (Baumeister, Vohs, & Funder, 2007). This decline can be seen to reflect a growing interest in models of psychopathology and intervention that offer perspectives beyond those afforded by behavioral analysis and learning theory. Broadly speaking, the focus in the literature has shifted from the environments that shape dysfunction to other forces that underpin it. Interestingly, however, the more that the neurosciences illuminate the role of biology in pathways to health and dysfunction, the more this research also highlights the very importance of the

environment. Some of the most compelling findings from such research relate to gene × environment interactions, in which genetic vulnerabilities confer risk for adverse outcomes only when combined with specific contexts (see Moffitt, Caspit, & Rutter, 2006). Prominent examples include Caspi and colleagues' (2002) finding that a functional polymorphism of the gene that encodes the neurotransmitter-metabolizing enzyme monoamine oxidase A (MAOA) moderated the effects of child maltreatment on risk for antisocial behavior in a longitudinal cohort. Low levels of MAOA expression were associated with significant risk for conduct disorder, but only among children who had been exposed to maltreatment early in life. Research into epigenetic processes has also attracted much attention, suggesting that environmental conditions in early life can structurally alter DNA and in turn produce risk for psychopathology over the life of the individual (Meaney, 2010).

Evidence of this kind is increasingly informing conceptualizations of risk and protection in relation to contextual dynamics, and in turn presenting researchers with new methodological challenges. Here we focus on three such challenges and the potential for observational coding to address the issues they raise. The first challenge concerns the theory-driven measurement of contextual dynamics when testing predictions that emphasize the interaction of individual-level and environment-level factors. The second concerns the reliable measurement of theoretically important individual differences using methods that can be implemented across diverse settings and study designs. The third challenge concerns the translation of emerging models of psychopathology into clinical interventions.

Investigating Emerging Models of Psychopathology

There is growing evidence that individual differences associated with biologically based characteristics interact—and transact—with environmental factors to shape trajectories of risk and protection. Various forms of evidence (e.g., experimental, longitudinal, genetic) have informed models of such processes in relation to distinct forms of psychopathology, with scientific advances allowing for increasingly precise conceptualizations of critical child–environment dynamics. In the area of antisocial behavior, developmental models have been informed by particularly rapid progress of this kind (e.g., Hawes, Brennan, & Dadds, 2009). To test the emerging predictions from such models, researchers will be increasingly required to measure contextual variables and processes that may be difficult—if not impossible—to characterize through self-report methods. We believe that the use of observation in this capacity will play a major role in future research of this kind. As addressed throughout this chapter, observational methods often provide the most sensitive measures of contextual dynamics, and importantly, can be adapted with great flexibility for purposes of theory-driven measurement. In recent years we have seen a range of innovative studies conducted in this vein, some examples of which follows.

While adverse child-rearing environments characterized by severe maltreatment have been associated with atypical neurocognitive development in children, little is known about the effects of normative variations in the child-rearing environment. In a design that incorporated the observation of parent–adolescent interactions in laboratory-based tasks and magnetic resonance imaging of adolescent brain structure, Whittle and colleagues (2009) examined whether normative variations in maternal responses to adolescents' positive affective behavior were associated with characteristics of adolescents' affective neural circuitry. Parent and adolescent affect and verbal content were coded from a pleasant event-planning interaction and a conflictual problem-solving interaction. The extent to which mothers exhibited punishing responses to their adolescents' affective behavior, as coded from these interactions, was associated with orbitofrontal cortex and anterior cingulate cortex volumes in these adolescents (Whittle et al., 2009). In this study direct observation was key to characterizing subtle variations in maternal socialization and to demonstrating the importance of these common relationship dynamics to the neuroanatomic architecture of children's social, cognitive, and affective development.

Longitudinal research has shown that children of depressed mothers exhibit relatively poor cognitive, neuropsychological, social, and emotional skills across childhood and adolescence. Predictions regarding the mechanisms through which this risk is conferred have focused increasingly on the compromised capacities for depressed mothers to construct a growth-promoting environment for their infants, with the relational behavior of such mothers characterized by reduced sensitivity, restricted range of affective expression, and inconsistent support of the infant's budding engagement (Goodman & Gotlib, 1999). However, empirical investigations of such mechanisms have relied largely on animal studies

that allow for the manipulation of environmental conditions. Rodent studies have shown that manipulating rearing conditions to simulate depression (e.g., preventing maternal licking and grooming of pups) disrupts the development of the hypothalamic-pituitary-adrenal (HPA) stress management system in offspring (see review by Champagne, 2008).

Researchers who have begun to test the predictions from these animal models in humans have relied heavily on direct observation, using methodologies that typically integrate observational coding with neurobiological measures. Feldman and colleagues (2009), for example, investigated such predictions in 9-month-old infants of mothers with postnatal anxiety and depression. Observational measurement served multiple purposes in this study, indexing aspects of socialization as well as infant temperament. Maternal sensitivity and infant social engagement were coded from mother–infant play interactions in the home environment. Infant fear regulation was also microcoded from a structured paradigm adapted from the Laboratory Temperament Assessment Battery (Goldsmith & Rothbart, 1996). Data from various self-report measures were also collected, and infant cortisol was assayed from salivary measures to index stress (HPA axis) reactivity. Echoing findings from the animal literature, maternal sensitivity was meaningfully related to infant social engagement and stress reactivity, while maternal withdrawal was associated with infant fear regulation. The integration of observational measurement into designs of this kind represents a promising means to investigate a range of physiological support systems that may be compromised by prenatal and postpartum exposure to adverse conditions associated with parental psychopathology.

Davies and colleagues (2008) used a highly structured analogue observation task to investigate a somewhat related research question concerning the association between children's biologically based characteristics and their reactivity to interparental conflict. Children witnessed a live simulated conflict and resolution scenario between their parents. This involved mothers acting from a script involving a disagreement with fathers over the telephone. Mothers were instructed in advance to convey mild irritation, frustration, and anger toward their partner as they normally would at home. Salivary cortisol was collected from children at three points during the simulated conflict, and three dimensions of children's behavioral reactivity to the conflict were coded from video recordings of the procedure: distress (e.g., freezing—tense, motionless, or fixed in place), hostility (e.g., anger—facial expressions or posters reflecting anger), and involvement (e.g., inquiries about parent feelings or relationships—questions about the emotional state of the parent or quality of the interparental relationship [e.g., "Mom, are you okay?," "Is Dad mad?"]). Relative to other forms of behavioral reactivity, children's distress responses to interparental conflict were consistent and unique predictors of their cortisol reactivity to interparental conflict. Furthermore, observed distress was particularly predictive of greater cortisol reactivity when children's observed levels of involvement in the conflict were also high.

Finally, in a novel study of emotion in young children, Locke and colleagues (2009) used observation to index affective responses that are inappropriate to the contexts in which they occur, and examined the association between this affect and salivary cortisol level. To measure context-inappropriate affect, children were administered a variety of episodes from the Laboratory Temperament Assessment Battery (Lab-TAB) designed to elicit negative affect (e.g., inhibition during conversation with a stranger, anger or sadness during a disappointment paradigm) or pleasure (e.g., anticipating surprising their parent). Observers then coded the presence and peak intensity of anger (e.g., bodily anger or frustration, anger vocalizations) in 5-second intervals. Displays of anger that were inappropriate to context were found to predict low levels of basal cortisol. Importantly, this prediction was unique from that afforded by levels of anger that were context-appropriate. Such findings support the importance of examining contextual aspects of emotion when investigating its role in relation to broader processes of risk and protection, and the value of observation for this purpose.

Theory-Driven Measurement of Individual Differences

In addition to the theory-driven measurement of contextual variables, it is becoming increasingly important for researchers to be able to characterize participants on dimensions related to biologically based factors. It is now commonly accepted that most psychopathologies and many complex behaviors have genetic origins, and that there are multiple routes to the same behavioral phenotype (or behavioral symptoms). In between genes and behavior are endophenotypes—individual differences that form the causal links between genes and the overt expression of disorders (see Cannon & Keller, 2006).

For example, there is evidence to suggest that mentalizing—the intuitive ability to understand that other people have minds—may be an endophenotype of the social impairments in autism (Viding & Blakemore, 2007). Growing research is concerned with the identification of endophenotypes for various disorders, placing increasing emphasis on the reliable and practical measurement of such individual differences. In our own research we have found observation to be of particular value in measuring individual differences associated with putative subtypes of antisocial behavior differentially associated with callous-unemotional (CU) traits. There is now considerable evidence that conduct problems follow distinct developmental trajectories in children with high versus low levels of CU traits, involving somewhat causal processes (see Frick & Viding, 2009).

Data from our initial experimental studies—using emotion-recognition and eye-tracking paradigms—suggested that children with high levels of CU traits exhibit deficits in the extent to which they attend to the eye regions of faces (Dadds et al., 2006, 2008). To move beyond these computer-based tasks and investigate whether this failure to attend to the eyes of other people occurs in real-world social interactions, we relied heavily on observational measurement. In our first such study (Dadds et al., 2011a), we observed the parent–child interactions of children with clinic-referred conduct problems, in analogue scenarios involving "free play," a family-picture drawing task, and an "emotion talk" task in which parents and children discussed recent happy and sad events. Parent–child interactions were coded using global ratings of social engagement, talk, and warmth to contextualize rates of parent–child eye contact. Interval coding was then used to code rates of eye contact. As previous literature on coding eye contact in family interactions could not be located, intervals of various length were compared in pilot testing to achieve an acceptable balance between measurement sensitivity and observer demands. While levels of eye contact were found to be reciprocated in mother–son and father–son dyads, boys with high levels of CU traits showed consistent impairments in eye contact towards their parents. Interestingly, although CU traits were unrelated to the frequency of eye contacts initiated by mothers, fathers of high-CU boys exhibit reduced eye contact toward their sons (Dadds et al., 2011a).

Based on these findings, we postulated that a failure to attend to the eyes of attachment figures could drive cascading errors in the development of empathy and conscience in children with high levels of CU traits (Dadds et al., 2011a). As no established technology existed for investigating the processes by which these deficits may be shaped by parenting dynamics—and potentially shape such dynamics in return—we subsequently developed a novel paradigm for this purpose. The "love" task (Dadds et al., 2011b) was expressly designed to elucidate parent–child interactions that are sensitive to the emotion-processing deficits associated with CU traits. The task concentrates parent–child interactions into a short but emotionally intense encounter for which reciprocated eye gaze is fundamental. It was administered following approximately 30 minutes of parent–child play and conversation and was prompted by an experimenter with the following instructions: "I'm going to come back into the room to do one more game. Once I have gone, I'd like you to look [child's name] in the eyes and show him/her, in the way that feels most natural for you, that you love him/her."

Video recordings of the subsequent 90-second interaction were coded using global ratings of mother and child levels of comfort and genuineness during the interaction, verbal and physical expressions of affection, and eye contact—both initiated and rejected. Compared with controls, children with oppositional defiant disorder were found to reciprocate lower levels of affection from their mothers, and those with CU traits showed significantly lower levels of affection than the children lacking these traits. As predicted, the high-CU group showed uniquely low levels of eye contact toward their mothers (Dadds et al., 2011b). This paradigm appears to be a promising tool for characterizing children with high levels of CU traits. Importantly, as an observational paradigm it is able to provide data that are unaffected by report biases. With growing evidence that child CU traits and family environment are associated with interacting and bidirectional risk processes (e.g., Hawes Brennan, & Dadds, 2009; Hawes et al., 2011), this method is likely to have broad applications in future research.

Observation in Translational Research

As a consequence of the growing impact of the neurosciences on models of psychopathology, the need for translational research is growing likewise. Findings from emerging intervention research have shown that contextual dynamics can be critical to understanding the interplay between behavioral and biological variables in therapeutic change. The potential for observational coding to capture such dynamics in translational research designs has also

been demonstrated. O'Neal and colleagues (2010), for example, examined child behavior and cortisol response as long-term (16-month) outcome variables in a randomized controlled trial of a parenting intervention for preschoolers at risk for conduct problems. Structured parent–child interactions were observed both in the laboratory and the family home, and multiple parent–child constructs were coded using various systems. The frequency of discrete aggressive behaviors was microcoded using the DPICS, and various global observation systems (e.g., Home Observation for the Measurement of the Environment–Early Childhood Version; Caldwell & Bradley, 1984) were used to code dimensions related to parental warmth and engagement. Not only was the intervention shown to prevent the development of aggressive behavior, but cortisol data demonstrated that it also resulted in normalized stress responses. The effect of the intervention on aggression was found to be largely mediated by the intervention effect on cortisol response, but only among families characterized by low levels of observed warmth (O'Neal et al., 2010). In addition to demonstrating the value of integrating observational measurement into such a design, such findings underscore the importance of parental warmth for the development of the HPA axis during early childhood, and suggest that HPA axis function is amenable to change through interventions that modify social environments in this period.

Large-scale randomized controlled trial designs, however, are not the only way to translate emerging models of psychopathology into clinical practice. In fact, small-scale designs that allow for the intensive examination of change processes may be more likely to inform the early stages of intervention development. Such an approach has recently been recommended for the purpose of developing new interventions in the field of autism (Smith et al., 2007). Just as early stimulus control studies used repeated observational measurement to translate operant theory into behavioral interventions for child conduct problems in single-case experimental designs (e.g., Patterson, 1974), we believe that such designs are now needed to translate conceptualizations of heterogeneous causal processes in emerging models of antisocial behavior (see Frick & Viding, 2009). Importantly, researchers testing the clinical application of novel theories are likely to be required to design novel, theory-driven observational strategies, as opposed to relying on "off-the-shelf" coding systems. In our own research we are currently conducting such investigations using the love task paradigm (Dadds et al., 2011b) to observe change in parent and child behaviors of theoretical importance to CU traits (e.g., eye contact) in response to novel interventions.

Summary

The potential for any research strategy to produce meaningful findings will be determined first and foremost by the meaningfulness of the research question, and as we have reviewed in this chapter, observational coding has proven to be well suited to a range of research questions in the clinical literature. Such coding has been widely used to index the dimensions of diagnostic symptoms associated with various disorders, the contextual dynamics of functional importance to these disorders, and individual differences (e.g., child temperament) and internal processes (e.g., cognitive biases) implicated in pathways to these disorders. In recent years considerable progress has been achieved in establishing high-quality coding systems for research with specific clinical populations—most notably discordant couples and the distressed families of children with conduct problems. At the same time, research involving the theory-driven adaptation of such systems, and the development of novel observational paradigms, has demonstrated that the flexibility associated with observational measurement remains one of its major strengths. The type of structure that is applied to elicit behavior in either analogue or naturalistic observation, as well as the methods by which this behavior is coded, recorded, and analyzed, can all be adapted for theoretical purposes. We believe that observational coding will be an important tool in emerging translational research, allowing researchers to operationalize various biologically based individual differences and capture critical information about the contexts in which they emerge.

References

Aksan, N., Kochanska, G., & Ortmann, M. R. (2006). Mutually responsive orientation between parents and their young children: Toward methodological advances in the science of relationships. *Developmental Psychology*, *42*, 833–848.

Allen, J. P., Insabella, G., Porter, M. R., Smith, F. D., Land, D., & Phillips, N. (2006). A social-interactional model of the development of depressive symptoms in adolescence. *Journal of Consulting and Clinical Psychology*, *74*, 55–65.

Aspland, H., & Gardner, F. (2003). Observational measures of parent child interaction. *Child and Adolescent Mental Health*, *8*, 136–144.

Bakeman, R., & Gottman, J.M. (1987). Applying observational methods: A systematic view. In J. Osofsky (Ed.), *Handbook of infant development* (2nd ed., pp. 818–854). New York: Wiley.

Barkley, R. A. (1997). Attention-deficit/hyperactivity disorder. In E. J. Marsh & L. G. Terdal (Eds.), *Assessment of childhood disorders* (3rd ed., pp. 71–129). New York: Guilford Press.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396–403.

Bellack, A. S., & Hersen, M. (1998). *Behavioral assessment: A practical handbook*. Elmsford, NY: Pergamon Press.

Birchler, G. R., Weiss, R. L., & Vincent, J. P. (1975). Multimethod analysis of social reinforcement exchange between maritally distressed and nondistressed spouse and stranger dyads. *Journal of Personality and Social Psychology, 31*(2), 349–360.

Brown, W. H., Pfeiffer, K. A., McIver, K. L., Dowda, M., Addy, C. L., & Pate, R. R. (2009). Social and environmental factors associated with preschoolers' nonsedentary physical activity. *Child Development, 80*, 45–58.

Caldwell, B. M., & Bradley, R. H. (1984). *Home Observation for Measurement of the Environment–Revised Edition*. Little Rock: University of Arkansas at Little Rock.

Cannon, T. D., & Keller, M. C. (2006). Endophenotypes in the genetic analyses of mental disorders. *Annual Review of Clinical Psychology, 2*, 267–290.

Caspi, A. (2000). The child is father of the man: Personality continuities from childhood to adulthood. *Journal of Personality and Social Psychology, 78*, 158–172.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., & Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, *297*, 851–853.

Champagne, F. A. (2008). Epigenetic mechanisms and the transgenerational effects of maternal care. *Frontiers in Neuroendocrinology, 29*, 386–397.

Cone, J. (1999). Observational assessment: Measure development and research issues. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 183–223). New York: Wiley.

Dadds, M. R., Allen, J. L., Oliver, B. R., Faulkner, N., Legge, K., Moul, C., Woolgar, M., & Scott, S. (2011b). Love, eye contact, and the developmental origins of empathy versus psychopathy. *British Journal of Psychiatry*, *198*, 1–6.

Dadds, M. R., Barrett, P. M., Rapee, R. M., & Ryan, S. (1996). Family processes and child anxiety and aggression: An observational analysis. *Journal of Abnormal Child Psychology, 24*, 715–734.

Dadds, M. R., El Masry, Y., Wimalaweera, S., & Guastella, A. J. (2008). Reduced eye gaze explains "fear blindness" in childhood psychopathic traits. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 455–463.

Dadds, M. R., Jambrak, J., Pasalich, D., Hawes, D. J., & Brennan, J. (2011a). Impaired attention to the eyes of attachment figures and the developmental origins of psychopathy. *Journal of Child Psychology and Psychiatry, 52*(3), 238–245.

Dadds, M. R., Perry, Y., Hawes, D. J., Merz, S., Riddell, A., Haines, D., Solak, E., & Dadds, M. R., & Hawes, D. J . (2006). *Integrated family intervention for child conduct problems*. Brisbane, Queensland: Australian Academic Press.

Dadds, M. R., & Sanders, M. R. (1992). Family interaction and child psychopathology: A comparison of two methods of assessing family interaction. *Journal of Child and Family Studies, 1*, 371–392.

Davies, P. T., et al. (2008). Adrenocortical underpinnings of children's psychological reactivity to interparental conflict. *Child Development, 79*(6), 1693–1706.

Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupéré, V. (2011). Improving classroom attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science, 12*, 70–79.

Dishion, T. J., & Andrews, D. W. (1995). Preventing escalation in problem behaviors with high-risk young adolescents: Immediate and 1-year outcomes. *Journal of Consulting and Clinical Psychology, 63*, 538–548.

Dishion, T. J., & Bullock, B. (2001). Parenting and adolescent problem behavior: An ecological analysis of the nurturance hypothesis. In J. G. Borkowski, S. Ramey, & M. Bristol-Power (Eds.), *Parenting and the child's world: Influences on intellectual, academic, and social-emotional development* (pp. 231–249). Mahwah, NJ: Erlbaum.

Dishion, T. J., & Granic, I. (2004). Naturalistic observation of relationship processes. In S. N. Haynes & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment* (Vol. 3): Behavioral assessment (pp. 143–161). New York: Wiley.

Dishion, T. J., Nelson, S. E., Winter, C., & Bullock, B. (2004). Adolescent friendship as a dynamic system: Entropy and deviance in the etiology and course of male antisocial behavior. *Journal of Abnormal Child Psychology, 32*, 651–663.

Dishion, T. J., & Snyder, J. (2004). An introduction to the "Special Issue on advances in process and dynamic system analysis of social interaction and the development of antisocial behavior." *Journal of Abnormal Child Psychology, 32*(6), 575–578.

Dishion, T., & Stormshak, E. (2007). *Intervening in children's lives: An ecological, family-centered approach to mental health care*. Washington, DC: American Psychological Association.

Dishion, T. J., & Tipsord, J. M. (2011). Peer contagion in child and adolescent social and emotional development. *Annual Review of Psychology, 62,* 189–214.

Dobson, K. S., & Kendall, P. C. (Eds.) (1993). *Psychopathology and cognition*. San Diego: Academic Press.

Driver, J. L., & Gottman, J. M. (2004). Daily marital interactions and positive affect during marital conflict among newlywed couples. *Family Process, 43*(3), 301–314.

Du Rocher Schudlich, T., & Cummings, E. M. (2007). Parental dysphoria and children's adjustment: marital conflict styles, children's emotional security, and parenting as mediators of risk. *Journal of Abnormal Child Psychology, 35*, 627–639.

Dubi, K., Emerton, J., Rapee, R., & Schniering, C.2008. Maternal modelling and the acquisition of fear and avoidance in toddlers: Influence of stimulus preparedness and temperament. *Journal of Abnormal Child Psychology, 36*, 499–512.

Ehrmantrout, N., Allen, N. B., Leve, C., Davis, B., & Sheeber, L. (2011). Adolescent recognition of parental affect: Influence of depressive symptoms. *Journal of Abnormal Psychology, 120*(3), 628–634.

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based treatments for child and adolescent disruptive behavior disorders. *Journal of Clinical Child and Adolescent Psychology, 37*, 213–235.

Eyberg, S., Nelson, M., Duke, M., & Boggs, S. (2004). *Manual for the Dyadic Parent–Child Interaction Coding System* (3rd ed.). Unpublished manuscript, University of Florida, Gainesville.

Feldman, R., Granat, A. D. I., Pariente, C., et al. (2009). Maternal depression and anxiety across the postpartum year and infant social engagement, fear regulation, and stress reactivity. *Journal of the American Academy of Child and Adolescent Psychiatry, 48*(9), 919–927.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria

for efficacy, effectiveness, and dissemination. *Prevention Science, 3*, 151–175.

Frick, P. J., & Viding, E. (2009). Antisocial behavior from a developmental psychopathology perspective. *Development and Psychopathology, 21*, 1111–1131.

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*, 99–118.

Gardner, F. (2000). Methodological issues in the direct observation of parent-child interaction: do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review, 3*, 185.

Gardner, F., Hutchings, J., Bywater, T., & Whitaker, C. (2010). Who benefits and how does it work? Moderators and mediators of outcome in an effectiveness trial of a parenting intervention. *Journal of Clinical Child & Adolescent Psychology, 39*(4), 568–580.

Goldsmith, H. H., & Rothbart, M. K. (1996). *The Laboratory Temperament Assessment Battery (LAB-TAB): Locomotor Version 3.0. Technical Manual*. Madison, WI: Department of Psychology, University of Wisconsin.

Goodman, S. H., & Gotlib, I. H. (1999). Risk for psychopathology in the children of depressed mothers: a developmental model for understanding mechanisms of transmission. *Psychological Review, 106*, 458–490.

Gottman, J. (1979). *Marital interaction: Experimental investigations*. New York: Academic Press.

Gottman, J. (1991). Chaos and regulated change in families: A metaphor for the study of transitions. In P. A. Cowan, & M. Heatherington (Eds.), *Family transitions* (pp. 247–372). Hillsdale, NJ: Erlbaum.

Gottman, J. M. (1998). Psychology and the study of marital processes. *Annual Review of Psychology, 49*, 169–197.

Gottman, J. M., Coan, J., Carrere, S., & Swanson, C. (1998). Predicting marital happiness and stability from newlywed interactions. *Journal of Marriage and the Family, 60*, 5–22.

Gottman, J. M., Guralnick, M. J., Wilson, B., Swanson, C. C., & Murray, J. D. (1997). What should be the focus of emotion regulation in children? A nonlinear dynamic mathematical model of children's peer interaction in groups. *Development and Psychopathology, 9*, 421–452.

Granic, I., & Lamey, A. V. (2002). Combining dynamic systems and multivariate analyses to compare the mother-child interactions of externalizing subtypes. *Journal of Abnormal Child Psychology, 30*(3), 265–283.

Granic, I., O'Hara, A., Pepler, D., & Lewis, M. D. (2007). A dynamic systems analysis of parent-child changes associated with successful "real-world" interventions with aggressive children. *Journal of Abnormal Child Psychology*, *35*, 845–857.

Green, S. B., & Alverson, L. G. (1978). A comparison of indirect measures for long-duration behaviors. *Journal of Applied Behavior Analysis, 11*, 530.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 109–138). New York: Plenum Press.

Hawes, D. J., Brennan, J., & Dadds, M. R. (2009). Cortisol, callous-unemotional traits, and antisocial behavior. *Current Opinion in Psychiatry*, *22*, 357–362.

Hawes, D. J., & Dadds, M. R. (2005a). Oppositional and conduct problems. In J. Hudson & R. Rapee (Eds.), *Current thinking on psychopathology and the family* (pp. 73–91). New York: Elsevier.

Hawes, D. J., & Dadds, M. R. (2005b). The treatment of conduct problems in children with callous-unemotional traits. *Journal of Consulting and Clinical Psychology, 73*(4), 737–741.

Hawes, D. J., & Dadds, M. R. (2006). Assessing parenting practices through parent-report and direct observation during parent-training. *Journal of Child and Family Studies, 15*(5), 555–568.

Hawes, D. J., Dadds, M. R., Frost, A. D. J., & Hasking, P. A. (2011). Do childhood callous-unemotional traits drive change in parenting practices? *Journal of Clinical Child and Adolescent Psychology*, *52*, 1308–1315.

Heyman, R. E. & Slep, A. M. S. (2004). Analogue behavioral observation. In M. Hersen (Ed.) & E. M. Heiby & S. N. Haynes (Vol. Eds.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 162–180). New York: Wiley.

Hollenstein, T., Granic, I., Stoolmiller, M., & Snyder, J. (2004). Rigidity in parent–child interactions and the development of externalizing and internalizing behavior in early childhood. *Journal of Abnormal Child Psychology, 32*, 595–607.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the LIFE coding system. *Journal of Clinical Child Psychology, 24*(2), 193–203.

Hudson, J., Comer, J. S., & Kendall, P. C. (2008). Parental responses to positive and negative emotions in anxious and non-anxious children. *Journal of Clinical Child and Adolescent Psychology, 37*, 1–11.

Hudson, J. L., Doyle, A., & Gar, N. S. (2009). Child and maternal influence on parenting behavior in clinically anxious children. *Journal of Clinical Child and Adolescent Psychology, 38*(2), 256–262.

Jacob, T., Tennenbaum, D. L., & Krahn, G. (1987). Factors influencing the reliability and validity of observation data. In T. Jacob (Ed.), *Family interaction and psychopathology: Theories, methods, and findings* (pp. 297–328). New York: Plenum Press.

Jacob, T., Tennenbaum, D., Seilhamer, R. A., Bargiel, K., & Sharon, T. (1994). Reactivity effects during naturalistic observations of distressed and nondistressed families. *Journal of Family Psychology*, *8*, 354–363.

Johnson, S. M., & Bolstard, O. D. (1975). Reactivity to home observation: A comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavior Analysis, 8*, 181–185.

Johnston, J. J., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Kerig, P. K., & Baucom, D. H. (Eds.). (2004). *Couple observational coding systems*. Mahwah, NJ: Erlbaum.

Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36*, 220–232.

Lewis, M. D., Lamey, A. V., & Douglas, L. (1999). A new dynamic systems method for the analysis of early socioemotional development. *Developmental Science, 2*, 458–476.

Locke, R. L., Davidson, R. J., Kalin, N. H., & Goldsmith, H. H. (2009). Children's context inappropriate anger and salivary cortisol. *Developmental Psychology, 45*(5), 1284–1297.

Lorber, M. F., O'Leary, S. G., & Kendziora, K. T. (2003). Mothers' overreactive discipline and their encoding and appraisals of toddler behavior. *Journal of Abnormal Child Psychology, 31*, 485–494.

Maerov, S. L., Brummet, B., & Reid, J. B. (1978). Procedures for training observers. In J. B. Reid (Ed.), *A social learning approach to family intervention: Vol. 11. Observation in home settings* (pp. 37–42). Eugene, OR: Castalia Press.

Mash, E. J., & Terdal, L. G. (1997). *Assessment of childhood disorders* (3rd ed.). New York: Guilford Press.

McNeil, C. B., & Hembree-Kigin, T. L. (2010). *Parent–child interaction therapy* (2nd ed.). New York: Springer.

Meaney, M. J. (2010). Epigenetics and the biological definition of gene × environment interactions. *Child Development, 81*, 41–79.

Merrilees, C. E., Goeke-Morey, M. C., & Cummings, E. M. (2008). Do event-contingent diaries about marital conflict change marital interactions? *Behavior Research and Therapy, 46*, 253–262.

Moffitt, T. E., Caspi, A., & Rutter, M. (2006). Measured gene-environment interactions in psychopathology: Concepts, research strategies, and implications for research, intervention, and public understanding of genetics. *Perspectives on Psychological Science, 1*, 5–27.

Moustakas, C. E., Sigel, I. E., & Schalock, M. D. (1956). An objective method for the measurement and analysis of child-adult interaction. *Child Development, 27*, 109–134.

O'Neal, C. R., Brotman, L. M., Huang, K., Gouley, K. K., Kamboukos, D., Calzada, E. J., et al. (2010). Understanding relations among early family environment, cortisol response, and child aggression via a prevention experiment. *Child Development, 81*, 290–305.

O'Rourke, J. F. (1963). Field and laboratory: The decision-making behavior of family groups in two experimental conditions. *Sociometry, 26*, 422–435.

Ost, L. G., Svensson, L., Hellstrom, K., & Lindwall, R. (2001). One-session treatment of specific phobias in youths: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 69*, 814–824.

Patterson, G. F. (1974). A basis for identifying stimuli which control behaviors in natural settings. *Child Development, 45*, 900–911.

Patterson, G. R. (1982). *Coercive family processes*. Eugene, OR: Castalia.

Patterson, G. R., & Chamberlain, P. (1994). A functional analysis of resistance during parent training therapy. *Clinical Psychology: Science and Practice, 1*, 53–70.

Patterson, G. R., Forgatch, M. S., & DeGarmo, D. S. (2010). Cascading effects following intervention. *Development and Psychopathology, 22*(4), 949–970.

Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys*. Eugene, OR: Castalia.

Piehler, T. F., & Dishion, T. J. (2007). Interpersonal dynamics within adolescent friendship: Dyadic mutuality, deviant talk, and patterns of antisocial behavior. *Child Development, 78*(5), 1611–1624.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis, 10*, 325–332.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology, 77*, 302–316.

Ryan, K. D., Gottman, J. M., Murray, J. D., Carrere, S., & Swanson, C. (2000). Theoretical and mathematical modeling of marriage. In M. D. Lewis & I. Granic (Eds.), *Emotion, development and self-organization: Dynamic systems approaches to emotional development* (pp. 349–372). New York: Cambridge University Press.

Schneider, B. H. (2009). An observational study of the interactions of socially withdrawn/anxious early adolescents and their friends. *Journal of Child Psychology and Psychiatry, 50*, 799–806.

Sheeber, L. B., Davis, B., Leve, C., Hops, H., & Tildesley, E. (2007). Adolescents' relationships with their mothers and fathers: Associations with depressive disorder and subdiagnostic symptomatology. *Journal of Abnormal Psychology, 116*, 144–154.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). The assessment of parenting practices in families of elementary school-aged children. *Journal of Clinical Child Psychology, 25*, 317–327.

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., et al. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders, 37*, 354–366.

Snyder, J., Reid, J. B., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science, 7*, 43–56.

Snyder, J., Schrepferman, L., McEachern, A., Barner, S., Provines, J., & Johnson, K. (2008). Peer deviancy training and peer coercion-rejection: Dual processes associated with early onset conduct problem. *Child Development, 79*, 252–268.

Stoolmiller, M., Eddy, J. M., & Reid, J. B. (2000). Detecting and describing preventative intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology, 68*, 296–305.

Trentacosta, C. J., Hyde, L. W., Shaw, D. S., Dishion, T. J., Gardner, F., & Wilson, M. (2008). The relations among cumulative risk, parenting, and behavior problems during early childhood. *Journal of Child Psychology and Psychiatry, 49,* 1211–1219.

Viding, E., & Blakemore, S-J. (2007). Endophenotype approach to the study of developmental disorders: implications for autism research. *Behavior Genetics, 37*, 51–60.

Whittle, S., Yap, M. B., Yucel, M., Sheeber, L., Simmons, J. G., & Pantelis, C., et al. (2009). Maternal responses to adolescent positive affect are associated with adolescents' reward neuroanatomy. *Social Cognitive & Affective Neuroscience, 4*(3), 247–256.

Zangwill, W. M., & Kniskern, J. R. (1982). Comparison of problem families in the clinic and at home. *Behavior Therapy, 13*, 145–152.

CHAPTER

9 Designing, Conducting, and Evaluating Therapy Process Research

Bryce D. McLeod, Nadia Islam, *and* Emily Wheat

Abstract

Therapy process research investigates what happens in therapy sessions and how these interactions influence outcomes. Therapy process research employs an array of methodologies but has recently used clinical trials as a platform for investigating process–outcome relations. This chapter serves as a resource for performing and interpreting therapy process research conducted within clinical trials. Issues related to designing, conducting, and evaluating therapy process research are reviewed, with examples drawn from the child therapy literature to illustrate key concepts. The chapter concludes with suggested future research directions.

Key Words: Alliance, therapeutic interventions, treatment integrity, therapy process, outcome

Therapy process research investigates what happens in psychotherapy sessions and how these activities influence clinical outcomes (Hill & Lambert, 2004). Process research covers many topics and employs diverse methodologies, with current efforts using clinical trials as a platform for process research (e.g., to investigate process–outcome relations). Randomized clinical trials (RCTs) can be an ideal vehicle for process research (Weersing & Weisz, 2002a). Collecting process data during an RCT can greatly increase the scientific yield of a clinical trial. Indeed, secondary data analysis of clinical trial data has played a role in identifying how evidence-based treatments (EBTs) produce change (e.g., Crits-Christoph, Gibbons, Hamilton, Ring-Kurtz, & Gallop, 2011; Huey, Henggeler, Brondino, & Pickrel, 2000), the relation of client involvement and outcome (Chu & Kendall, 2004; Coady, 1991; Edelman & Chambless, 1993, 1994), whether or not therapeutic tasks affect the alliance (Kendall, Comer, Marker, Creed, Puliafico, et al., 2009), and the strength of the alliance–outcome association (Chiu, McLeod, Har, & Wood, 2009; Hogue, Dauber, Stambaugh, Cecero, & Liddle, 2006; Klein et al., 2003). The results of these studies can facilitate the dissemination and implementation of EBTs into community settings (Kendall & Beidas, 2007; McLeod & Islam, 2011; McLeod, Southam-Gerow, & Weisz, 2009).

The goal of this chapter is to serve as a resource for those conducting and interpreting therapy process data collected within an RCT, with examples drawn from the child therapy literature to illustrate key concepts. Issues related to designing, conducting, and analyzing therapy process studies will take the forefront. Therapy process can, for example, include client behavior (e.g., developing social skills), therapist behavior (e.g., therapeutic interventions such as cognitive restructuring), and facets of the relation between client and therapist (e.g., level of client involvement; quality of the client–therapist alliance). Outcome refers to the short- and long-term changes in the client brought about by therapy (Doss, 2004).

Overview of Therapy Process Research

Before focusing on process research methods, consider a conceptual framework. Figure 9.1 depicts a model that incorporates theory and findings from the process research tradition (Doss, 2004) and

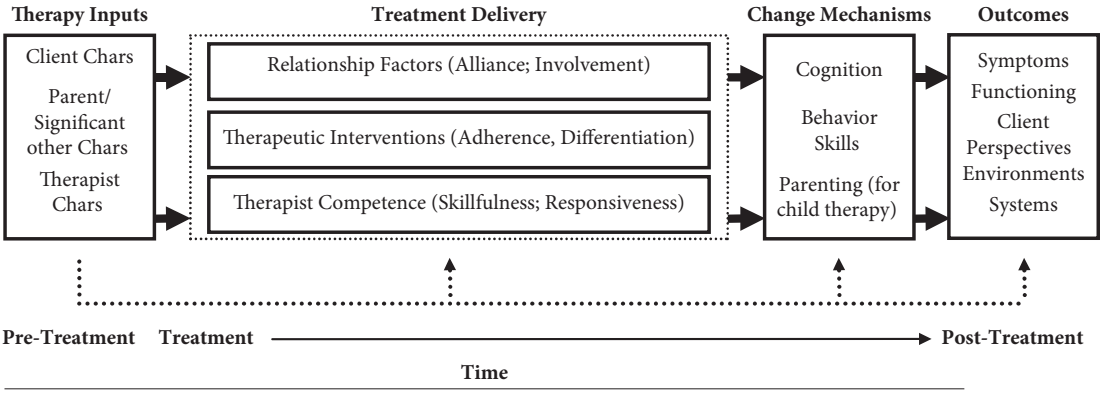


Figure 9.1 Theoretical Model of Therapeutic Change in Therapy.

treatment integrity research that investigates the degree to which EBTs are delivered as specified in treatment manuals (Dane & Schneider, 1998; Hogue, 2002; Jones, Clarke, & Power, 2008; Waltz, Addis, Koerner, & Jacobson, 1993). The model details how the three components of therapy process—client, therapist, and relational factors—affect clinical outcomes. Although developed for youth psychotherapy, the model can be extended to apply to therapy with participants of any age. An in-depth review regarding each facet of the model is beyond the scope of this chapter, but the model provides a framework to understand how the components discussed may together, or in isolation, influence outcomes.

Psychotherapy Inputs

The left side of the model identifies therapy inputs that may influence or moderate the process and outcome of therapy. Therapy inputs include (a) client characteristics, such as symptom severity (Ruma, Burke, & Thompson, 1996); (b) parent/significant other characteristics, such as psychopathology (Cobham, Dadds, & Spence, 1998); (c) family characteristics, such as stress and family income level (Kazdin, 1995); (d) therapist characteristics, such as theoretical orientation (Weersing, 2000) or attitudes toward manual-based treatments (Aarons, 2005; Becker, Zayfert, & Anderson, 2004); and (e) service characteristics, such as organizational culture and climate (Schoenwald, Carter, Chapman, & Sheidow, 2008). These inputs represent factors present at the start of treatment that potentially influence process and outcome.

Process Factors

The middle section depicts the main focus of this chapter, the core components involved in treatment delivery: therapeutic interventions (e.g., changing cognitive distortions), therapist competence, and relational factors (e.g., alliance, client involvement). Each component is hypothesized to facilitate symptom reduction (Chu & Kendall, 2004; Kendall & Ollendick, 2004; Orlinsky, Ronnestad, & Willutzki, 2004).

The delivery of specific *therapeutic interventions* is hypothesized to promote symptom reduction (e.g., McLeod & Weisz, 2004; Silverman, Pina, & Viswesvaran, 2008). An emerging area of focus that can aid understanding of how therapeutic interventions affect outcomes is treatment integrity research (McLeod et al., 2009). Treatment integrity focuses upon the degree to which a treatment is delivered *as intended* (Perepletchikova & Kazdin, 2005; Waltz et al., 1993). Two components of treatment integrity, treatment adherence and differentiation, refer specifically to the type of therapeutic interventions delivered by the therapist. Treatment adherence refers to the extent to which the therapist delivers the treatment as designed (e.g., delivers the prescribed interventions contained within a treatment manual). Treatment differentiation refers to the extent to which a therapist delivers therapeutic interventions proscribed by a specific treatment manual (e.g., delivering psychodynamic interpretations in a cognitive-behavioral treatment [CBT] program). These two treatment integrity components therefore identify the prescribed (and proscribed) therapeutic interventions that together, and/or in isolation, are hypothesized to be responsible for change (Perepletchikova & Kazdin, 2005).

Therapist competence, a second component of treatment integrity (Perepletchikova & Kazdin, 2005), is key to treatment delivery (Kazdin & Kendall, 1998). Competence refers to the level of

skill and degree of responsiveness demonstrated by a therapist when delivering the technical and relational elements of therapy (Perepletchikova & Kazdin, 2005; Waltz et al., 1993). A therapist's ability to deliver interventions with skill and responsiveness is said to maximize their effects. To date, research has revealed mixed findings regarding the strength of the relation between therapist competence and outcomes (Webb, DeRubeis, & Barber, 2010). Perhaps, in studies where most or all therapists meet a standard of implementation and are monitored, there is little variability and therefore limited association with outcome.

Relational factors—both the alliance and client involvement—have been found to be related to symptom reduction (Braswell, Kendall, Braith, Carey, & Vye, 1985; Chu & Kendall, 2004; Horvath & Bedi, 2002; Manne, Winkel, Zaider, Rubin, Hernandez, & Bergman, 2010; McLeod, 2011). A therapist's abilities to (a) cultivate a relationship with the client (child, parent, adult, couple, family) marked by warmth and trust and (b) promote the client's participation in therapeutic activities are considered instrumental in promoting positive outcomes (Chu et al., 2004; Chu & Kendall, 2004; Horvath & Bedi, 2002). It has been hypothesized that a strong alliance facilitates positive outcomes via increased client involvement in therapeutic tasks (Kendall & Ollendick, 2004; Manne et al., 2010), although support for this hypothesis has been mixed (Karver et al., 2008; Shirk, Gudmundsen, Kaplinski, & McMakin, 2008). Compared to the adult field, the relational elements of therapy have received relatively little empirical attention in the youth field (McLeod, 2011).

Change Mechanisms

Change mechanisms represent the means through which therapy produces change (Doss, 2004; Kazdin, 1999). Using youth as an example, components that have been hypothesized as mechanisms of change include habituation (e.g., Bouchard, Mendlowitz, Coles, & Franklin, 2004; Hollon et al., 2002) and cognitive change (Bouchard et al., 2004; Kendall & Treadwell, 2007; Treadwell & Kendall, 1996). Other change mechanisms, such as problem solving, relaxation, and self-monitoring, are posited to produce change, although evidence is needed. Studying change mechanisms is important for learning how EBTs work (Weersing & Weisz, 2002a). Understanding what produces change can advance treatment, refine therapist training, and improve outcomes.

Psychotherapy Outcomes

The right portion of the diagram represents treatment outcomes. Hoagwood and colleagues (Hoagwood, Jensen, Petti, & Burns, 1996) suggested five outcome domains: (a) symptoms/diagnoses, a primary outcome in RCTs; (b) functioning, defined as the ability to meet the demands of home, work, peer group, or neighborhood; (c) consumer satisfaction, defined as the client's experience and/or satisfaction with the mental health services; (d) environments, changes in a specific aspect of the client's life (e.g., home, work) brought about by therapy (e.g., changes in family and/or couple communication); and (e) systems, assessment of service use patterns following treatment. For process research, although all domains are relevant, the reduction of symptoms and improvements in functioning are two key domains.

The model provides a framework for the factors that may influence the process and outcome of therapy. In addition, the model aids understanding of how process components may be studied in isolation (e.g., the alliance–outcome relation) and/or in combination (e.g., therapist competence and client involvement).

The Methods of Process Research

Broadly speaking, the research strategies can be divided into qualitative and quantitative methods. Qualitative approaches, such as having a therapist review a therapy session and comment upon specific processes, offer some desirable features. For example, a qualitative approach provides an opportunity to gather in-depth information from participants and hear their unique perspective and experience of therapy. However, qualitative approaches are time-intensive, are vulnerable to bias, and are not particularly well suited to examine the relation between process and outcome across clients in a RCT. Quantitative approaches, such as those used in integrity research (e.g., Carroll et al., 2000; Hogue et al., 2008), employ measures that produce numerical scores that can be used to describe and analyze therapy processes. The quantitative approach allows researchers to aggregate findings across participants (e.g., determine the mean competence level of therapists in a clinical trial), so it is particularly well suited for using RCTs as a platform for process research. For this reason, and in keeping with the systematic empirical focus of this Handbook, the current chapter focuses exclusively upon quantitative approaches.

Therapy Process Dimensions

Process research is inherently complex. It is therefore useful to have a system that classifies therapy process measures along critical dimensions. Such a system achieves at least two important goals. First, a classification system provides a tool to evaluate existing process measures. Second, classification helps researchers design new process measures. The framework presented below builds upon previous systems and represents the product of a tradition of therapy process measure categorization (c.f., Hill & Lambert, 2004).

Target

Therapy process measures typically target the client, therapist, or supervisor. Of course, within each category exists a variety of configurations, such as client (child, parent, adult), therapist (single therapist, co-therapists, treatment team), supervisor (single supervisor, supervision team), or combination (client–therapist, therapist–supervisor relationship). Researchers must decide whom to target (typically the client and the therapist), but this decision can become complex when therapy involves more than one person. In such cases, the researcher must decide whether to consider them as one or more targets. For example, some alliance measures for youth psychotherapy define the target as the child (e.g., Shirk & Saiz, 1992), whereas other alliance measures define the target as the child–therapist dyad (e.g., McLeod & Weisz, 2005). The decision on the target depends upon the research question and the intervention under study.

Focus

Therapy process measures focus upon four process domains: (a) behavior (client, therapist, supervisor), (b) thematic content, (c) style, and (d) quality. Client or therapist behavior is a common focus of therapy process measures, and this can be divided into overt (observable) and covert (nonobservable) categories. Overt behaviors include, for example, verbal (e.g., open-ended questions, reflections) or physical (e.g., behavioral involvement, eye contact) behaviors that can be directly observed. Covert behaviors cannot be directly observed and can, for example, include client or therapist intentions or level of motivation. Process measures can focus upon thematic content (e.g., family issues, peer relationships, substance use). Style represents another domain and is defined as the manner in which a client or therapist acts (e.g., condescendingly, empathetically, warmly, critically). Process measures that assess quality focus on the skill with which particular behaviors are performed (e.g., therapist competence; Perepletchikova & Kazdin, 2005).

Theoretical Foundation

All process measures are based, at least in part, upon a theoretical model. Some process measures are developed to assess core elements of a specific theoretical orientation. For example, the Cognitive Therapy Adherence and Competence Scale (Barber, Liese, & Abrams, 2003) assesses therapist adherence and competence in cognitive therapy. Other process measures are designed to assess what have been called "common factors" that have been hypothesized to promote positive outcomes across different theoretical orientations (Karver, Handelsman, Fields, & Bickman, 2005). For example, the Working Alliance Inventory (Horvath & Greenberg, 1989) assesses the working alliance, which is posited to be associated with positive outcomes across different therapies (Bordin, 1979). The conceptual foundation of a process measure has implications for scale development and, in part, determines its potential research applications. For example, the study of youth therapy has yet to coalesce around a single definition of the alliance. As a result, existing measures of alliance are designed to assess different theoretical conceptualizations. Unfortunately, this state of affairs makes comparing findings across studies difficult (McLeod, 2011). The theoretical foundation of a process measure needs to be considered when ascertaining whether it is a good fit for a particular research question.

Perspective

Measures of process variables rely upon multiple perspectives: the client, therapist, supervisor, and/or judge. Clients and therapists are participating observers as they are involved in the therapy sessions. Due to this involvement, they offer a unique perspective about events that occur in therapy, including interactions that may not be recorded. In contrast, judges do not directly participate in the therapy process and make their ratings from recordings or transcripts of sessions. Although judges may be considered objective, there are forces that may bias their ratings (Hill & Lambert, 2004). Although entirely removing bias is unlikely, we later discuss steps that can be taken to minimize potential sources of bias for both participants and judges.

Unit of Measurement

Process measures can be placed into two broad categories of measurement: microprocess and macroprocess. Microprocess measures focus upon small units (e.g., utterances, single words, speaking turns) of a therapy session. Given this specificity, microprocess measures are typically assessed using judges. Macroprocess measures have a global focus (e.g., therapy session, course of treatment). This distinction among process measures is consistent with Hawes, Dadds, and Pasalich's distinction between microsocial and macrosocial observational coding methods (see Chapter 8 in this volume). Most process measures used in RCTs focus upon macroprocess, because this unit of measurement is seen as appropriate for evaluating process–outcome relations (Hogue, Liddle, & Rowe, 1996).

Theoretical and methodological matters need to be considered when determining the unit of measurement. When a theoretical model details how the process–outcome relation unfolds within treatment, this information can determine the unit of measurement. For example, a session or portion of a session may be needed to assess a process (e.g., client involvement; Chu & Kendall, 2004), whereas multiple sessions may be needed to assess other processes (e.g., transference; Bordin, Cutler, Dittmann, Harway, Rausch, & Rigler, 1954). Methodologically, investigators must decide whether or not to use predetermined units of measurement. With predetermined units, researchers specify when ratings are made (e.g., after a specific number of minutes or sessions). Without predetermined units, judges determine on a case-by-case basis when to make ratings. Predetermined units increase interrater reliability (necessary for data analysis), but may, in the opinion of some (Marmar, 1990), restrict information. Always consider the impact that the unit of measurement may have on ratings generated by specific process measures (Hill & Lambert, 2004). For a more thorough consideration of interval-based time sampling methods in observational research, see Chapter 8 in this volume.

Type of Measurement

Process measures are generally of three types: nominal scales, interval scales, and Q-sort. Nominal systems involve placing process data into predetermined categories. These categories can be either mutually exclusive or not mutually exclusive (Hill & Lambert, 2004). Mutually exclusive systems require judges to categorize the same unit into one category. For example, the Helping Skills System (Hill & O'Brien, 1999) assigns one skill (e.g., open sentence, immediacy, direct guidance) to each therapist response unit (e.g., every grammatical sentence). A system that is not mutually exclusive allows judges to categorize the same unit into multiple categories. For example, Southam-Gerow and colleagues (2010) used an 11-item measure to assess therapist adherence to a CBT program for youth anxiety. For the measure, judges watched an entire therapy session and rated whether a therapist delivered one or more CBT interventions prescribed in the treatment manual. A single session may involve multiple CBT interventions, and thus several items can be rated. When using a system that is not mutually exclusive, it is important to distinguish between items that covary. Each item should be treated independently and rated as though it is completely uncorrelated with others. Compared to alternate measurement types, nominal systems can be more reliable because they focus exclusively upon the presence or absence of certain behaviors. Nominal ratings are often used for descriptive purposes because scores are not averaged across judges.

Interval rating scales typically involve rating processes on a Likert scale. Both even- and odd-numbered scales are used in interval rating systems (e.g., scales that vary from 5 to 10 points). Even-numbered scales are used when one wants to force judges to opt for positive or negative ratings. Examples of interval rating scales include the Experiencing Scale (Klein, Mathieu-Coughlan, & Kiesler, 1986), the Therapeutic Alliance Scale for Children (Shirk & Saiz, 1992), and the Therapy Process Observational Coding System for Child Psychotherapy Strategies scale (TPOCS-S; McLeod & Weisz, 2010). Interval rating scales are used because they provide average scores across judges for a given instance, an entire session, or the course of treatment. To average scores across judges, the data must meet the assumption of interval scales: that there is an equal difference between scale points and that some points on the scale are "better" than others.

The Q-sort has judges sort items along a rating scale with a forced distribution. This method requires judges to rate items in relation to one another and necessitates that the full range of the scale is used. One example of this method is the Process Q-Set (Jones, Cumming, & Horowitz, 1988), which describes the interaction between the client and therapist using items related to therapist and client attitudes and behaviors. Judges indicate which items best describe a therapy session using a normal distribution of 5, 8, 12, 16, 18, 16, 12, 8,

and 5 items in each of nine categories respectively (1 = least characteristic, 5 = neutral, 9 = most characteristic). The drawback of the Q-sort method is that it forces a particular distribution on the items and can present difficulties when selecting methods of statistical analysis.

Level of Inference

Therapy process measures vary in the level of inference required to produce ratings. Noninferential measures entail little interpretation of states or intentions, such as a process measure that requires judges to make ratings based upon overt behaviors (e.g., client eye contact or posture). Inferential measures ask judges to discern a speaker's intentions or internal states based on observation (e.g., transference). Because ratings for noninferential measures are based on overt behaviors, interrater reliability is typically better (Hill & Lambert, 2004). In contrast, inferential measures may require more experienced judges and have lower interrater reliability because the ratings typically depend upon the judge's interpretation of observed behaviors.

Stimulus Materials

Judges rely upon a variety of stimuli to produce process ratings, including therapy sessions, transcripts, audio recordings, video recordings, or some combination thereof. Therapy sessions are commonly used as stimuli. Therapists and clients may make retrospective ratings following a session, or judges may make ratings while watching a session recording. Most studies rely upon audio or video recordings (Hill, Nutt, & Jackson, 1994). This is because recorded stimuli provide the most accurate account of what happens in a session. Recording sessions may, however, influence client and therapist behavior (cf. Hill & Lambert, 2004). Ideally, researchers should use a combination of stimuli (e.g., recorded and transcribed material) to ensure accurate content.

Planning, Conducting, and Evaluating Therapy Process Research

Next, issues related to planning, conducting, and evaluating therapy process research are addressed. This section focuses upon the methodological and practical factors associated with conducting therapy process research.

Measure Selection and Evaluation

When designing a therapy process study, researchers must decide whether or not to use an existing measure. Typically, a literature search is conducted to identify measures that assess the construct of interest. If process measures are identified, then the researcher must decide whether or not to use one of the measures. This decision should be based in part upon (a) the design of the measure and whether it assesses the process dimensions relevant to the planned study and (b) the psychometric properties of the measure. There are several advantages to using an existing measure. These include allowing researchers to both compare findings across studies and amass data on the psychometric properties of a measure (McLeod, 2011).

However, existing measures do not always assess the construct of interest, in which case a researcher may decide to develop a new measure. Developing a new measure is time intensive and costly. Moreover, researchers who develop process measures do not always use them again, which can slow scientific progress by making it difficult to compare findings across studies (McLeod, 2011). Researchers must therefore carefully consider whether or not it is better to use an existing measure or develop a new one. See Tables 9.1 and 9.2 for a comparison of several of the more widely used and/or evaluated process measures across key dimensions and features.

Reliability and Validity of Therapy Process Measures

The potential value of a process measure depends, in part, upon its psychometric properties. In the following section, the major categories of reliability and validity that pertain to process research are reviewed.

RELIABILITY

Reliability assesses the consistency and dependability of a measure or of judges' ratings. There are a number of reliability dimensions along which a process measure or judges' ratings might be assessed. The following paragraphs provide an overview of the different forms of reliability and when they are typically assessed.

Broadly defined, a *reliable* measure is one that delivers consistent results and is free from measurement error. The reliability of a measure (i.e., internal consistency, test–retest) is reported when multi-item process measures rated on interval scales are used. Internal consistency evaluates the degree to which items on a scale assess the same construct (e.g., the alliance). According to Devellis (2003), Cronbach alpha coefficients of .65 or less are considered "minimally acceptable," coefficients greater

Table 9.1 Comparison of Key Dimensions and Features Across Widely Used and/or Evaluated Relational Measures

	Alliance									Involvement					
	CALPAS	Pennsylvania Scales	CPPS	SOFTA	TASC	TPOCS—A	Vanderbilt Scales	ATAS	WAI	CIRS	FEQ	OBBAHCM	ORTI	MCS	OAES
Process Factors															
Interventions															
Therapist competence															
Relational elements	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Process Dimensions															
Target															
Child			■	■	■	■	■	■	■	■	■			■	■
Parent				■	■	■		■	■						
Adult	■	■		■			■	■						■	
Therapist	■	■	■	■	■	■	■	■	■						
Combination(s)	■	■	■	■	■	■	■	■	■						
Focus															
Overt	■	■	■	■	■	■	■	■		■				■	■
Covert	■	■			■				■						
Style															
Quality															
Theory															
Generic	■		■	■	■	■		■	■		■			■	
Embedded in theory		■			■		■								■
Measurement															
Microprocess															
Macroprocess	■	■	■	■	■	■	■	■	■		■				
Nominal scale(s)															
Interval scale(s)	■	■	■	■	■	■	■	■	■		■				
Q-sort															
Evaluator															
Client	■	■	■		■				■					■	
Therapist	■	■	■		■				■			■		■	
Supervisor															
Judge	■	■	■	■		■	■	■	■		■				■

■ = Dimension/factor addressed in measure

Note: CALPAS = California Psychotherapy Alliance Scales (Marmar et al., 1989); Pennsylvania Scales (Luborsky, 1976); Child Psychotherapy Process Scales (Estrada & Russell, 1999); SOFTA = System for Observing Family Therapy Alliances (Friedlander et al., 2006); TPOCS-A = Therapy Process Observational Coding System for Child Psychotherapy—Alliance Scale (McLeod & Weisz, 2005); Vanderbilt Scales (Gomes-Schwartz, 1978); ATAS = Adolescent Therapeutic Alliance Scale (Johnson et al., 1998); WAI = Working Alliance Scale (Horvath & Greenberg, 1989); CIRS = Client Involvement Scale (Chu & Kendall, 2004); OBBAHMC = Observer-Based Behavioral Activation Homework Completion Measure (Busch, Uebelacker, Kalibatseva, & Miller, 2010); ORTI = Observational ratings of therapy involvement (Braswell, Kendall, Braith, Carey, & Vye, 1985); FEQ = Family Engagement Questionnaire (Kroll & Green, 1997); OAES = Overall Adolescent Engagement Scale (Jackson-Gilfort, Liddle, Tejeda, & Dakof, 2001).

Table 9.2 Comparison of Key Dimensions and Features Across Widely Used and/or Evaluated Intervention and Competence Measures

	CTS	CSPRS	CTACS	PPQ	SAM	TAM	TBRS	TBRS—Competence	TPC	TPOCS-S	YACS
Process Factors											
Interventions		■	■	■	■	■	■	■	■	■	■
Therapist competence	■	■	■					■			■
Relational elements											
Process Dimensions											
Target											
Child											
Parent											
Adult				■							
Therapist	■	■	■	■	■	■	■	■	■	■	■
Combination(s)											
Focus											
Overt	■	■	■	■	■	■	■	■	■	■	■
Covert											
Style											
Quality											
Theory											
Generic				■							
Embedded in theory	■	■	■		■	■	■	■	■	■	■
Measurement											
Microprocess											
Macroprocess	■	■	■		■	■	■	■	■	■	■
Nominal scale(s)											
Interval scale(s)											
Q-sort				■							
Evaluator											
Client						■					
Therapist					■				■		
Supervisor											
Judge	■	■	■	■			■	■		■	■

■ = Dimension/factor addressed in measure

Note: CTS = Cognitive Therapy Scale (Young & Beck, 1980); CSPRS = (Hollon et al., 1988); CTACS; PPQ; SAM = Supervisor Adherence Measure (Schoenwald et al., 1998); TAM = Therapist Adherence Measure (Henggeler & Borduin, 1992); TBRS = Therapist Behavior Rating Scale (Hogue et al., 1998); TBRS—Competence = Therapist Behavior Rating Scale—Competence (Hogue et al., 2008); TPC = Therapy Procedures Checklist (Weersing, Donenberg, & Weisz, 2002); TPOCS-S = Therapy Process Observational System for Child Psychotherapy—Strategies Scale (McLeod & Weisz, 2010); YACS = Yale Adherence and Competence Scale (Carroll et al., 2000).

than .65 are considered "respectable," and those .80 or more are considered "very good."

Test–retest reliability assesses the concordance of scores on a measure completed multiple times by the same reporter. This form of reliability is used to assess constructs that are believed to remain stable over time (e.g., intelligence). Pearson correlation coefficients are used to assess test–retest reliability and are considered acceptable if .70 or more. Test–retest reliability is rarely reported for process measures because most process variables are not expected to remain stable (Chu & Kendall, 2004; Hill & Lambert, 2004).

When judges are used in process research, the reliability of their observations must be reported. Intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979) are typically used to assess interrater reliability when judges use interval rating scales to produce process ratings. According to Cicchetti (1994), ICC values below .40 reflect "poor" agreement, from .40 to .59 reflect "fair" agreement, from .60 to .74 reflect "good" agreement, and .75 and higher reflect "excellent" agreement.

When using ICC estimates, researchers must select the appropriate model (Hill & Lambert, 2004; Shrout & Fleiss, 1979). Estimates can be produced for a single judge or the mean of a group of judges. When a single judge rates all recordings and reliability with another judge is calculated on a subset (e.g., 20 percent) of the recordings, then the appropriate ICC estimate is single rater. However, if all recordings are coded by two or more judges, then the correct ICC estimate is the mean of raters.

ICCs can also be based upon a fixed- or random-effects model. A fixed-effects ICC model provides a reliability estimate for the consistency of a single rater or the mean of a group of raters for a particular sample. That is, the reliability estimate does not generalize beyond the particular group of judges used in a specific study. A random-effects model provides a reliability estimate of a single rater or the mean of a group of raters and allows for generalizability of the results to other samples. If judges are randomly sampled from the general population, then a random-effects model is appropriate; however, if judges are not sampled randomly from the population, then a fixed-effects model is appropriate. Most studies utilize the random-effects ICC model because it is assumed that judges are randomly sampled from the population.

A useful framework that allows researchers to investigate interrater reliability, along with measure reliability and validity, is generalizability theory (Brennan, 2001). Generalizability theory is a statistical framework for investigating and designing reliable observations (e.g., alliance ratings, treatment integrity scores) across different conditions of measurement. This approach allows researchers to examine the performance of process measures across facets (i.e., sources of variation) relevant to different research applications of the process measure. For example, researchers can investigate whether therapists, clients, or treatment phase account for a significant amount of variation in treatment adherence ratings. If therapists account for a significant proportion of the variance, then this suggests that therapists differ significantly in treatment adherence. The generalizability framework therefore allows researchers to determine whether key facets systematically influence therapy process ratings. This information enables researchers to investigate the psychometric properties of a measure and use the resulting knowledge to design efficient measurement procedures capable of producing dependable observations (Brennan, 2001). For example, using a generalizability framework, Crits-Christoph and colleagues (Crits-Christoph et al., 2011) found that adequate assessment of alliance requires multiple clients per therapist and at least four sessions per client. This approach also permits researchers to approximate ICC estimates. Using this framework, variance due to extraneous factors can be partialed out to produce a more accurate estimate of interrater reliability.

Finally, when assessing nominal categories, the Kappa coefficient is often used. This coefficient is ideal for use with nominal categories because it controls for chance agreement across raters and guessing. Fleiss (1981b) suggests that a Kappa less than .40 is "unacceptable," between .40 and .75 is "fair" to "good," and above .75 is "strong" agreement.

VALIDITY

Validity refers to whether or not a measure assesses what it purports to measure. There is no agreed-upon single definition for validity, and the dimensions of validity relevant to a particular process measure vary depending upon what it is designed to measure. The following paragraphs cover different validity dimensions relevant for process measures and detail when they are important to establish.

Face validity and *content validity* are important to establish early in the measure development process before assessing reliability or other forms of validity. Measures that have face validity appear appropriate for the purpose for which they are used. Face validity is not considered a critical validity dimension

because it does not help establish whether a process measure assesses what it is designed to assess. That being said, face validity may be important when participants are asked to fill out self-report process measures (Hill & Lambert, 2004). In such cases, participants may question the purpose of filling out a measure if the items do not appear to assess the purported construct (e.g., alliance). Face validity can therefore help increase the acceptability of process measures, which could improve participant compliance and data accuracy.

Content validity means that items capture all facets of a given construct. When developing a new measure, it is important for researchers to carefully document the steps taken to ensure content validity. To establish content validity, researchers can sample items from a wide range of measures designed to assess the same or similar construct. Researchers can also rely upon experts. Once a list of items is generated, experts can review item definitions, use the measure, and provide feedback that can be used to refine existing items or add new ones.

Measures achieving *construct validity* show evidence that obtained scores reflect the theoretical concept that the measure was designed to assess (Foster & Cone, 1995; Hill & Lambert, 2004). Construct validity is established by demonstrating that a process measure covaries with another measure in a predicted pattern above and beyond what can be ascribed to shared method variance (DeVellis, 2003). Establishing construct validity is accomplished over time and through numerous studies. The main forms of construct validity relevant to process research are convergent validity, discriminant validity, and predictive validity.

Convergent validity is achieved by showing that a process measure has a strong relation to another measure examining the same construct. For example, a significant correlation between scores on the Therapy Process Observational Coding System for Child Psychotherapy—Alliance (TPOCS-A) scale and the Therapeutic Alliance Scale for Children supported the convergent validity of the measures (Fjermestad, McLeod, Heiervang, Havik, Ost, & Haugland, 2012; McLeod & Weisz, 2005). Discriminant validity is achieved by showing that a process measure has low correlations with measures that assess unrelated constructs. For example, scores on the TPOCS-A were not related to treatment credibility (Fjermestad et al., in press), which supports the discriminant validity of the TPOCS-A. Finally, predictive validity is achieved when a process measure is associated with an expected outcome. For example, demonstrating that family-focused interventions, as measured by the Therapist Behavior Rating Scale (TBRS), for adolescent substance abuse predicted reductions in clinical outcomes supported the predictive validity of the TBRS (Hogue, Liddle, Dauber, & Samuolis, 2004).

Collecting Therapy Process Data

This section covers the procedural details associated with conducting a process study. To ensure data integrity, researchers must carefully attend to the procedures for collecting process data and maintain control over all aspects of data collection. Also, it is important to minimize sources of systematic bias that might influence study findings.

WORKING WITH RECORDINGS

Researchers who plan to record sessions need to take steps to ensure the quality of the recordings and completeness of the data. Prior to data collection, researchers must decide how to record therapy sessions. It is possible to make different types of recordings (audio, video) in various formats (VHS, DVD, digital). Audio recordings are inexpensive and less intrusive, but certain information is not captured (e.g., nonverbal behavior). In contrast, video recordings are more costly, but this format captures the most information. It is important to consider how recordings will be stored. Certain formats are easier to store, copy, and access across research sites. For example, digital files located on encrypted servers allow researchers from multiple sites to access recordings without having to copy the original recording. Digital files can therefore save time and money if recordings need to be shared. Finally, to minimize missing data, researchers should be responsible for recording sessions and use a system to track when a session occurred and whether it was recorded.

WORKING WITH THERAPISTS AND CLIENTS

Participants are key collaborators in the research process. Collecting self-report data from participants can provide access to covert behaviors (e.g., beliefs, attitudes) that are not accessible via other methods. Despite this strength, researchers need to take steps to ensure the quality of the process data by minimizing potential sources of bias.

In youth psychotherapy, the developmental level and reading ability of child participants need to be considered when selecting process measures. Children, adolescents, and adults differ significantly along important developmental factors

(e.g., cognitive and linguistic development) that can influence ratings. Many process measures used in youth therapy represent downward extensions of adult measures, so the reading level or item wording may not be appropriate for youth. Researchers must therefore assess whether particular process measures are appropriate for youth participants.

Therapist ratings on process measures may be influenced by their level of experience. Trainees may be more critical of their performance, whereas more experienced therapists may have the perspective and breadth of knowledge to provide more nuanced critiques (Hill & Lambert, 2004). Experienced therapists may provide more accurate ratings of complex processes, such as therapist competence (Waltz et al., 1993). However, if a therapist develops hypotheses based upon a particular theoretical orientation, biased ratings may result, especially if a process measure is based upon an alternate theoretical orientation (McLeod, 2009).

Prior to data collection, it is important to form a positive research relationship with participants by communicating that they play an important role in the research project. To maintain a positive research relationship, participating in research should not place an undue burden upon participants. Having to spend too much time completing measures and procedures can deter clients and therapists from participating. Moreover, the more time and effort a study requires, the less generalizable it is to other therapy situations.

When participant-rated process measures require training, researchers should take steps to ensure that participants understand how to use the measure before data collection begins. The training period should provide an opportunity to practice using the process measure before the start of data collection. A practice period helps to eliminate a "break-in" period during data collection in which participants provide inaccurate or incomplete data because they are learning to use the measure.

Finally, it is vital to emphasize the confidentiality of responses and maintain the privacy of participants. Clients may be biased toward reporting positively, especially if they believe that their therapist might view their responses. Similarly, therapists may be biased toward reporting positively if they believe that a supervisor may view their responses. To guard against this effect, it is important to have measures completed in a separate room and placed in a box. Participants should remain blind to study hypotheses.

Sampling Issues

Researchers must decide upon a sampling plan. Theoretical or empirical findings may inform the sampling plan, but this information does not always exist. In such cases, a sampling plan must address two issues. Researchers must first decide whether or not to sample within a therapy session (e.g., make multiple process ratings within a single session) or make ratings based upon an entire therapy session. Previous research has demonstrated that therapist and client behavior can vary within a therapy session (see, e.g., O'Farrell, Hill, & Patton, 1986). If a therapy process fluctuates within a session, then process data should be sampled multiple times within a session. However, if a process is consistent across a session, then a portion of the session can be sampled.

Researchers must also decide when to sample data across treatment. To determine how to sample data across treatment, three questions must be answered. (1) Is there a theoretical or methodological reason for sampling a particular stage of treatment? For example, the quality of the alliance and clinical outcomes may become confounded as treatment progresses, so it is preferred to assess alliance early in treatment (e.g., first three sessions; Kazdin, 2007). (2) How many sessions need to be sampled in order to produce a representative sample of treatment? Ideally, this question should be determined by data or informed by theory. Sampling too many sessions can be costly, whereas sampling too few sessions can produce unreliable or inaccurate data (Crits-Christoph et al., 2011). (3) Is it important to capture the trajectory of a therapy process variable over the course of treatment? A positive alliance trajectory across treatment is hypothesized to relate to positive clinical outcomes (Liber et al., 2010). To evaluate this hypothesis, cases can be divided into "early," "middle," and "late" treatment stages so the trajectory of the process variable can be assessed over time.

Ultimately, researchers must choose a sampling plan that fits their research question. If a researcher is studying a process that occurs consistently throughout a session or across treatment, then a smaller portion of therapy may be sampled. If a researcher is studying a variable that changes within a session or occurs infrequently (e.g., crying, self-disclosure), then more data points will need to be sampled.

Judges

A number of process measures rely upon trained judges to produce ratings. When working with

judges, researchers must first decide how many judges are needed. The project timeline may, in part, determine the number of judges. However, the expected reliability of the process measure should also be considered. More judges are needed when a low reliability is expected, as pooling the ratings from multiple judges can increase reliability.

Researchers must also select judges. It is generally assumed that characteristics of judges influence process ratings, but little research has evaluated this hypothesis (Hill & Lambert, 2004). Without empirical evidence to guide the selection process, researchers must decide if particular skills may be needed to make accurate ratings on a measure. One important factor to consider when selecting judges is the level of inference required for process ratings. Undergraduates may be a good fit for coding systems that require little inference; however, graduate students or experienced clinicians may be needed for coding systems that require more inference (e.g., therapist competence; Waltz et al., 1993).

Once selected, judges must be trained. Trainers should be coding experts who can meet regularly with the judges. Initially, training should include discussion of readings about the target construct, review of the coding manual, and review of exemplar recordings that illustrate key points from the coding manual. As training progresses, judges should engage in independent coding of recordings that represent the different facets of the target construct. Typically, judges first learn to identify the presence of a particular variable (e.g., an exposure) and then learn to calibrate their ratings with an established criterion (e.g., high, medium, and low ratings on an adherence scale). Ideally, training should expose judges to recordings that contain the complete range of scores on the measure, as this helps reduce the likelihood that range restriction will be a problem during coding. To complete training, judges should meet a specific criterion of interjudge agreement for each item (i.e., ICC > .60) with expert-generated ratings. This approach is preferred because it increases the likelihood of producing replicable judgments across trainers and sites, although it is not always possible to utilize this approach (e.g., when piloting a coding system). Throughout the training period and beyond, judges should remain blind to hypotheses related to the measure.

Once judges are trained, researchers must assign recordings. When judges do not code the entire sample, it is important to protect against bias by distributing the error associated with each judge across the sample. To achieve this goal, researchers typically use a randomized incomplete block design (Fleiss, 1981a) to assign each judge to the same number of sessions across relevant facets of the study (e.g., treatments, therapists).

Once coding begins, it is important to monitor for coder drift, in which ratings start out reliable following training but begin to shift as coding progresses. To protect against drift, researchers should hold regular coder meetings in which item definitions are discussed and specific recordings reviewed. In addition, researchers should perform regular interrater reliability checks (e.g., every 2 to 4 weeks) to check for coder drift. When interrater reliability for particular items (a) decreases for three or more reliability checks or (b) drops below a certain threshold (ICC = .60), then judges need to be retrained.

Data Analytic Considerations

In this section, issues related to preparing process data for analysis and different options for analyzing process data are reviewed.

Preparing Process Data for Analysis

Prior to testing study hypotheses, therapy process data need to be prepared for analysis. This process begins with an inspection of the data that includes a review of the distributional properties of each variable and checks for outliers. A nonnormal distribution, or the presence of outliers, may indicate data entry errors or the need to use transformed variables or nonparametric tests.

Once the preliminary inspection is complete, reliability is typically evaluated. This step includes assessing the reliability of the process measure and/or calculation of interrater reliability. If reliability is low, steps must be taken to address the problem. For example, items on observer-rated process measures are sometimes dropped when the ICC is below .40, unless it can be demonstrated that the item is theoretically important (e.g., Hogue et al., 2004) or inclusion of the item does not have a negative impact on scale reliability (e.g., Hogue, Dauber, Samuolis, & Liddle, 2006). The steps taken to address low measure and/or interrater reliability need to be carefully documented so subsequent researchers can assess the psychometric properties of a given measure.

Finally, researchers must decide how to operationalize each process variable. Investigators must first decide how a score for each measure, or subscale, will be generated from individual items

(e.g., averaged, summed, or combined in some other form). Researchers must then determine how to model the process variable within a session and/or across treatment. A couple of methodological issues, such as how to address clustering of data, must be considered when combining process data to generate scores. These issues will be covered in a subsequent section.

Descriptive Approaches

Descriptive approaches produce valuable information about specific therapy processes. Typically used in the discovery phase of research, descriptive approaches help define the nature of a therapy process. Frequency counts, proportions, and Likert-type ratings can all be used to describe a therapy process. Examples of recent descriptive studies are found in efforts to characterize the treatment provided to youth and their families in community-based service settings, called usual clinical care (UC). Effectiveness research often evaluates how an EBT delivered in a practice setting compares to the outcomes produced in UC. In some cases, EBTs have not outperformed UC (e.g., Southam-Gerow et al., 2010; Weisz et al., 2009), which has led to the question of what interventions are typically delivered in practice settings. The TPOCS-S (McLeod & Weisz, 2010) was developed to characterize UC and has been used to describe UC for youth with internalizing (McLeod & Weisz, 2010) and disruptive behavior (Garland et al., 2010) problems. Not only does each study serve as an example of descriptive process research, but the studies also represent a new generation of research that attempts to use process methods to aid efforts to improve the quality of mental health care in practice settings (Garland, Hurlburt, & Hawley, 2006).

Benchmarking is another promising descriptive tool. Benchmarking studies evaluate whether therapist performance in community settings approximates the performance standards achieved by therapists in efficacy trials. To date, benchmarking studies have primarily focused upon treatment outcomes observed in non-RCT studies in community settings (e.g., Wade, Treat, & Stuart, 1998; Weersing & Weisz, 2002b). However, benchmarking methods can also be used to study therapist behavior, such as treatment adherence and competence. Research suggests that community therapists do not deliver the full dose of EBTs in effectiveness trials (e.g., Southam-Gerow et al., 2010; Weisz et al., 2009); however, it is not known whether community therapists approach performance standards achieved by therapists in efficacy trials. For example, benchmarking analyses might consist of comparing treatment adherence and competence data collected in a community setting (e.g., an effectiveness trial) with treatment adherence and competence data from the same EBT delivered in a laboratory-based setting (e.g., an efficacy trial). Determining whether community therapists do, or do not, approximate the adherence, competence, and dosage standards achieved by their laboratory-based counterparts has important implications for implementation research and represents another application of descriptive analytic approaches.

Correlational Approaches

The correlational approach is commonly used in process research to relate a process variable (e.g., frequency count, extensiveness rating) to another process variable (e.g., alliance to client involvement; see Shirk et al., 2008) or an outcome. For example, a number of studies have evaluated the relation between the alliance and outcome in youth therapy employing a correlational approach (e.g., McLeod & Weisz, 2005; Shirk et al., 2008). Correlational approaches have a number of advantages; however, researchers must take steps to avoid certain limitations of this approach.

When using a correlational approach, particular attention must be paid to how the process variable is scored. Certain scoring strategies, such as frequency counts, are not a good match for a correlational approach. For example, when studying therapeutic interventions, the exclusive use of frequency counts can misrepresent the therapeutic process by giving a higher weight to interventions that are used more often, but not in a more thorough manner (Greenberg, 1986). As a result, scoring strategies that assess both the breadth and depth of a therapy process are more appropriate for correlational approaches (Hogue et al., 1996). An example of a scoring strategy that is appropriate for a correlational approach is called "extensiveness" ratings (Carroll et al., 2000; Hill, O'Grady, & Elkin, 1992; Hogue et al., 1996). Extensiveness ratings consider the breadth and depth of intervention delivery when generating scores. Such systems account for contextual factors, whereas frequency counts do not, and are therefore better suited for correlational approaches.

Hierarchical Approaches

Hierarchical approaches are ideal for analyzing process data for a number of reasons. First, in clinical

trials it is common for clients to be nested within therapists and therapists to be nested within sites. Standard analytic models, such as those used in correlational research, are generally not interpretable when applied to nested designs because the error terms of units (a therapist or site) are typically correlated within each level of analysis (Zucker, 1990). Therefore, analytic approaches that are appropriate for dealing with the nesting of clients within therapists need to be employed (see Barber et al., 2006; Carroll et al., 2000; Schoenwald, Sheidow, Letourneau, & Liao, 2003). Second, hierarchical approaches can deal with varying numbers of sessions per clients and varying numbers of clients per therapist (see, e.g., Hawley & Weisz, 2005). These challenges cannot be accommodated with standard analytic methods. Third, hierarchical approaches are ideal for assessing the temporal relation of process and outcome variables over the course of treatment (e.g., latent difference score models; see Crits-Christoph et al., 2011; Teachman, Marker, & Clerkin, 2010; Teachman, Marker, & Smith-Janik, 2008).

Measuring Outcomes

In process studies, researchers must pay careful attention to how and when clinical outcomes are assessed. In a clinical trial, it is typical to assess outcomes at pretreatment and posttreatment; however, this approach is not ideal for process research. For process research, both process and outcome should be assessed at the same time over the course of treatment, as this approach helps address temporal sequencing of the process and outcome variables (Weersing & Weisz, 2002a). For example, assessing the alliance and outcome during the same session early in treatment allows researchers to ascertain whether a significant alliance–outcome relation can be explained by prior symptom improvement (Klein et al., 2003). In addition to the repeated assessment of clinical outcomes, it is important to use dimensional measures designed for repeated assessment (Doss, 2004). Latent growth curve modeling strategies can be used to evaluate the rate and shape of change across key process variables and outcome measures across time (see Chapter 16 in this volume for comprehensive consideration of such analytic approaches).

Design Issues and Considerations

When conducting process research, investigators must anticipate and address a few issues that can affect the interpretability of study findings. These issues need to be considered when developing a study, collecting the data, and analyzing the findings.

Nesting

Nesting of clients within therapists and therapists within practice sites commonly occurs with process data. A nested data structure has the potential to influence study findings, as nesting can create dependencies within the data (Wampold & Serlin, 2000). That is, clients seen at the same clinic might have outcomes that are more highly correlated with one another than with outcomes of clients seen at another clinic. Standard analytic models, such as the fixed-effect general linear model, are generally not interpretable when applied to nested designs because the error terms of units are typically correlated within each level of analysis (Zucker, 1990). In such cases, standard analytic models can result in inflated type I error rates (Wampold & Serlin, 2000). Researchers must therefore decide how to deal with potential dependencies in the data.

There are several ways researchers can deal with nesting. In the planning phase, researchers need to determine if nesting will exist and, if so, ascertain how participant recruitment will be affected. To produce stable estimates, methodologists recommend that a minimum of six to eight clients need to be nested within each therapist (Hedeker, Gibbons, & Flay, 1994; Norton, Bieler, Ennett, & Zarkin, 1996). In some situations, getting six clients nested within each therapist may prove difficult, so researchers should take this into consideration when developing a sampling plan. Researchers must also account for the potential effect of nesting upon study power. To do so, the proposed sample size for the study must be adjusted to account for potential nesting. For example, the sample size can be multiplied by an inflation factor that takes into account the average cluster size and projected ICC (Donner, Birkett, & Buck, 1981).

Once data collection is finished, researchers can check for dependencies in the data by calculating ICCs (Norton et al., 1996). If dependencies exist in the data (ICC > .25; see Guo, 2005), then analytic approaches appropriate for dealing with the nesting of clients within therapists need to be employed (see Barber et al., 2006; Carroll et al., 2000; Schoenwald et al., 2003). If a study does not have enough clients nested within therapists, then investigators can treat the nested structure of the data as an extraneous variable that is controlled (see, e.g., Hogue et al., 2004, 2006, 2008).

Therapist Main Effects

Another design issue that warrants attention is *therapist main effects*. This term refers to potential differences among therapists in terms of therapy process (e.g., level of alliance) or outcome variables (Crits-Christoph & Mintz, 1991). In other words, certain therapists may consistently demonstrate higher (or lower) levels of competence or produce better (or worse) outcomes. It therefore is recommended to investigate whether mean-level differences exist between therapists on both process and outcome variables (Crits-Christoph & Mintz, 1991; Hogue et al., 2006). If systematic differences are identified, then this effect must be accounted for in subsequent analyses to ensure that the findings can be generalized to other samples.

Causality and the Direction of Effects

A common critique of process research is that the findings do not provide clear implications about causality or the direction of effects in the relation between therapy process and outcomes. To assert that a causal relation between a process and outcome variable exists, three conditions must be met (Feeley, DeRubeis, & Gelfand, 1999; Judd & Kenney, 1981): (a) the process and outcome variables must covary; (b) "third" variables must be ruled out (i.e., nonspuriousness); and (c) the process variable must precede the outcome variable. These three conditions are not typically met in most process research (Kazdin, 2007). Many process measures are collected at the end of treatment, making it impossible to establish the direction of effects. And many studies fail to rule out "third" variables (e.g., therapist experience) that may account for the process–outcome association. However, researchers can avoid these critiques by taking steps to meet the three conditions.

Responsiveness Critique

When conducting process–outcome research, it is important for investigators to understand the issues raised by the responsiveness critique (Stiles & Shapiro, 1989, 1994). This critique states that there are problems with applying the "drug metaphor"—that more of a process component translates to better outcomes—to process–outcome research due to the interactive nature of psychotherapy. Rather than acting in a random fashion, therapists match the delivery of process components to the client's level of functioning (Stiles & Shapiro, 1989, 1994). This means that there is a bidirectional relation between therapist and client behavior, as opposed to the unidirectional relation posited by the drug metaphor. According to this critique, if a therapist is perfectly responsive to a client's level of functioning, then the correlation between the frequency of particular therapist behaviors and improvements in client outcomes would be zero. Moreover, if a therapist delivers specific interventions more often to clients with more severe symptomatology, then the correlation between the frequency of the intervention and client outcomes would be negative. Negative correlations between process and outcome that have run counter to the hypothesized effect have been observed in a some studies (Castonguay, Goldfried, Wiser, Raue, & Hayes, 1996; Piper, Azim, Joyce, & McCallum, 1991). Process researchers employing correlational, regression, structural equation modeling, and ANOVA-based models must therefore be aware of the responsiveness critique.

The way in which therapy process is assessed may help address some concerns raised by the responsiveness critique (Doss, 2004). Stiles and Shapiro focused upon the use of frequency counts (e.g., number of interpretations, number of exposures). Frequency counts are subject to the responsiveness critique because they do not take context into account. Other approaches to rating therapy process, such as the previously discussed extensiveness ratings (see Hogue et al., 1996) or competence ratings (see Hogue et al., 2008), take context into consideration and thus are not as susceptible to this critique.

State of Knowledge: Examples from the Youth Psychotherapy Field

The past decade has witnessed calls for more therapy process research in youth psychotherapy. Specifically, researchers have called for more research focused upon the alliance (Kendall & Ollendick, 2004; Shirk & Karver, 2003), treatment integrity (Perepletchikova & Kazdin, 2005), and other processes related to youth outcomes (e.g., client involvement; Chu & Kendall, 2004). Investigators have heeded these calls and begun to expand knowledge in key areas. In this section, we focus upon two areas of process research that have received increased attention. A description of the recent advances is provided along with a discussion of what work remains to be done.

Alliance

Researchers and clinicians agree that the alliance represents an important ingredient of successful psychotherapy for youth. However, alliance research

in youth psychotherapy has lagged far behind the adult psychotherapy field. Whereas hundreds of alliance studies have been completed in the adult field (Horvath & Bedi, 2002), only 10 studies were completed in the youth field by 2006 (Karver, Handelsman, Fields, & Bickman, 2006). As a result, important questions about the nature and strength of the alliance–outcome relation in youth therapy exist.

To address questions about the alliance–outcome association, a recent meta-analysis synthesized the literature in the youth psychotherapy field (McLeod, 2011). The study set comprised 38 studies and the weighted mean effect size estimate of the alliance–outcome association (ES) was $r = .14$. This effect size estimate was smaller than those generated by previous meta-analyses focused upon the alliance–outcome association in adult and youth psychotherapy (r's > .20; Horvath & Symonds, 1991; Karver et al., 2006; Martin, Garske, & Davis, 2000; Shirk & Karver, 2003).

At first blush, these findings suggest that the alliance may explain a small proportion of the variance in clinical outcomes. However, it may be premature to draw this conclusion. Although the youth alliance literature has grown considerably in the past 5 years, it is still relatively small. Furthermore, the meta-analysis identified both theoretical (i.e., child age, problem type, referral source, mode of treatment) and methodological (i.e., source and timing of alliance assessment; domain, technology, and source of outcome assessment; single vs. multiple informants) moderators of the alliance–outcome association. This suggests that the small collection of existing studies do not represent a homogeneous collection and that more research is needed to address existing measurement and methodological issues.

It is possible to draw parallels between the status of alliance research and the challenges that face process research in the youth field. A number of methodological and theoretical issues will require attention for the field to advance. To illustrate some of the issues facing process researchers, we will note some of the prominent issues in the alliance field. First, few alliance measures are used across multiple studies. Across the 38 studies included in the meta-analysis, only five alliance measures (two observational and three self-report) were used more than once (McLeod, 2011). In fact, 16 distinct measures focused upon the child–therapist alliance were used (McLeod, 2011). As a result, study-to-study differences exist in how the alliance is conceptualized and measured. This variability makes it difficult to compare findings across studies and serves as a cautionary tale for researchers. Although it is sometimes necessary to develop new process measures, doing so can lead to a proliferation of measures that can slow scientific progress.

Second, few alliance measures have amassed reliability and validity data across multiple studies. It therefore is difficult to ascertain whether alliance measures tap into the same construct (Elvins & Green, 2008). To determine if alliance measures conceptually overlap, studies that assess the convergent validity of existing measures are needed (Elvins & Green, 2008; McLeod, 2011). There are few measures in the therapy process field that have well-established psychometric properties. Well-designed clinical trials that collect multiple process measures (e.g., child-, parent-, therapist-, observer-rated alliance measures) could help address this issue by investigating the validity of process measures.

Third, beyond methodological considerations, the field would benefit from research that addresses theoretical issues in the field. No study in the youth alliance field has included design elements that would help establish a causal relation between alliance and outcome. Moreover, only a handful of studies have evaluated whether the alliance exerts an influence on clinical outcomes via other treatment processes such as client involvement (see Karver et al., 2008; Shirk et al., 2008, for notable exceptions). This issue, in fact, is representative of a larger problem in process research in youth psychotherapy—that is, a lack of theory specification and testing. Further theory specification that details the relation between particular therapy processes and outcomes is needed to advance knowledge in the field.

Treatment Integrity Research

Treatment integrity research represents an exciting area of process research that has recently garnered increased attention in the youth psychotherapy field. To draw valid inferences from clinical trials, treatments must be well specified, well tested, and carried out as intended (Kazdin, 1994). Treatment integrity refers to the degree to which a treatment was delivered as intended and comprises three components—treatment adherence, treatment differentiation, and therapist competence (Perepletchikova & Kazdin, 2005; Waltz et al., 1993). As noted previously, treatment adherence refers to the extent to which the therapist delivers the treatment as designed. Treatment differentiation refers to the extent to which treatments under study differ along appropriate lines defined by the

treatment manual. Therapist competence refers to the level of skill and degree of responsiveness demonstrated by the therapist when delivering the technical and relational elements of treatment. Each component captures a unique aspect of treatment integrity that together, and/or in isolation, may be responsible for therapeutic change or lack thereof (Perepletchikova & Kazdin, 2005).

Treatment integrity research has pioneered the development of process measures that are uniquely suited to investigate process–outcome relations. In this research, observational assessment represents the gold standard because it provides objective and highly specific information regarding clinician within-session performance (Hill, 1991; Hogue et al., 1996; Mowbray, Holter, Teague, & Bybee, 2003). Indeed, observational assessment that incorporates four design elements produces process data with the maximum degree of reliability, validity, and utility (Carroll et al., 2000; Hogue, 2002; Waltz et al., 1993): (a) quantitative measures are used to investigate the intensity and frequency of interventions (i.e., extensiveness); (b) both model-specific interventions (therapeutic interventions that are essential to the underlying clinical theory) and common elements (therapeutic elements endorsed by most models, such as the alliance) are targeted; (c) quality (competence) as well as quantity (adherence) is assessed; and (d) both therapist (e.g., adherence) and client contributions (e.g., involvement) are considered. These design elements produce observational measures that are ideally suited to assess process–outcome relations (Hogue et al., 1996) and thus are uniquely suited to efforts designed to refine and optimize EBTs.

Although treatment integrity research is underdeveloped in the youth field, recent work illustrates the potential of this research to inform efforts to transport EBTs to practice settings. As noted previously, the TPOCS-S was designed to provide a means of objectively describing UC for youth (McLeod & Weisz, 2010). Recently, the TPOCS-S was used to characterize the treatment provided in a clinical trial evaluating the effectiveness of a CBT program for youth depression relative to UC. Clinicians employed by the community clinics were randomly assigned to provide either CBT (with training and supervision) or UC. At posttreatment, groups did not differ significantly on symptom or diagnostic outcomes, which raised questions about the effectiveness of CBT for youth depression. To enhance the informational value of the clinical trial, the TPOCS-S was used to address two questions relevant to interpreting the findings: (1) What interventions were used in UC? and (2) Were the CBT and UC conditions distinct (treatment differentiation)? Findings indicated that UC therapists used a wide range of interventions from multiple theoretical orientations but generally favored nonbehavioral approaches (e.g., client-centered interventions such as providing positive regard). The CBT and UC conditions were distinct as CBT sessions scored higher than UC on CBT interventions (e.g., problem solving). However, the CBT sessions received relatively low scores on CBT interventions ($M = 2.75$ on 7-point scale). This suggests that the therapists may have delivered a relatively low dose of CBT, which may explain why CBT did not outperform UC (Weisz et al., 2009).

Using the TPOCS-S to characterize the treatment provided in the effectiveness trial illustrates how treatment integrity research can play an important role in implementation research. Indeed, therapy process research provides researchers with the tools to document whether (and how) the delivery of EBTs changes when delivered in practice settings. Measures that capture critical aspects of treatment delivery, such as the TPOCS-S, can therefore help researchers interpret findings generated by effectiveness research and enhance their informational value. Although the TPOCS-S has a number of strengths, the measure does not capture all facets of treatment delivery (e.g., therapist competence). Thus, more work is needed to develop measures designed to capture all aspects of treatment integrity.

Future Directions

The therapy process field has recently made laudable advances; however, it is also clear that certain limitations that characterize the extant literature need to be addressed, particularly in the area of youth therapy process, as the previous examples illustrate. Questions raised about issues of causality require further refinement and evaluation of our theoretical models. Moreover, attention toward establishing the psychometric properties of existing process measures is needed. In this section, we discuss future directions for the field that can build upon existing strengths and address existing gaps.

Measurement Issues

Before the potential of therapy process research can be fully realized, attention must be paid to measure development and validation. Measures that assess key processes such as treatment integrity, alliance, and client involvement with demonstrated

psychometric properties are needed to move the field forward. Translating the need for psychometrically sound measures to reality will take a concerted effort and careful planning. RCTs represent an ideal platform for this research.

As research progresses, it will also be important for the field to expand upon the number of measurement options available to researchers. Observational assessment represents the gold standard in therapy process research (Hill, 1991; Hogue et al., 1996; Mowbray et al., 2003). However, despite its advantages, observational coding is time and resource intensive (Hill, 1991). Not all researchers have the resources to carry out observational coding (Schoenwald, Henggeler, Brondino, & Rowland, 2000; Weersing, Weisz, & Donenberg, 2002), so the development of supervisor-, therapist-, child-, and caregiver-report measures represents an important goal for the field (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Mihalic, 2004; NIMH, 1999). The potential of this approach is exemplified by research conducted with multisystemic therapy (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 1998), as parent reports of adherence to multisystemic therapy have been linked to outcomes (Henggeler, Melton, Brondino, Scherer, & Hanley, 1997; Henggeler, Pickrel, & Brondino, 1999; Huey et al., 2000). Developing self-report measures therefore represents an important goal for the field.

Addressing Issues of Causality

The issue of causality represents a consistent critique leveled at the therapy process field. To evaluate the competing explanations of the association between therapy process and outcomes, the field needs to move from basic correlational research to methodologies that can reveal more about the direction of causality. No study to date has (a) tested therapy process as a causal influence on outcomes, (b) ruled out the possibility that change in outcomes causally affects therapy process, or (c) ruled out that a third variable (e.g., client characteristics, such as symptom severity) affects both therapy process and outcomes systematically (or that there is an additive or multiplicative combination of more than one of these possibilities).

Two suggestions are offered to help the field begin to address issues of causality. First, because a specific sequence of events is a necessary, but not sufficient, precondition to establishing causality, utilizing repeated measures of therapy process and outcomes at meaningful time intervals will help test for the sequencing order implicit in causal models. Repeated measures could help test the temporal requirements of a causal relationship between systematic changes of therapy process and changes in clinical outcomes. Ideally, such research designs would employ multiple (e.g., measured at each session) points of data for both process and outcome and employ statistical modeling (e.g., hierarchical linear modeling) of the change in the trajectory of therapy process on the change in the trajectory of clinical outcomes, and vice versa.

Second, experimental methods can be employed to directly evaluate the effects of manipulating therapy process on outcomes, and vice versa. Intervention designs can help clarify the direction of effects between therapy process and clinical outcomes. In an intervention design participants are randomly assigned to either a condition that alters a therapy process or a condition that does not (e.g., a dismantling study that isolates particular therapeutic interventions). Therapy process and clinical outcomes are measured before and after the intervention. Strong evidence for the therapy process influencing outcomes exists if (a) the process intervention improves the outcome measure more than does the control intervention; (b) the therapy process improves clinical outcomes interactions more than does the control intervention; and (c) improvements in the outcome measure are mediated by improvements in the therapy process. Of course, even results that meet these three conditions do not "prove" that the therapy process causes a given outcome. However, such results provide more convincing evidence that the therapy process could have a causal effect.

Further Specification of Theories and Hypotheses

Greater theoretical specificity is needed to guide research on the role that specific therapy processes may play in promoting positive clinical outcomes, as well as to inform intervention programs. Broad models, such as the one presented in Figure 9.1, are a useful starting point for conceptualizing the multiple pathways through which therapy processes may contribute to clinical outcomes. However, there are few theory-derived hypotheses proposed in the literature about how specific therapy processes might operate in unison, or as part of a temporal chain, to affect clinical outcomes in youth psychotherapy. To advance, it will be necessary to specify the temporal processes involved and posit whether the expected effect would be on other therapy processes (i.e.,

alliance influences client involvement), short-term outcomes (i.e., skill acquisition during CBT), or long-term outcomes. For example, it is hypothesized that a strong child–therapist alliance facilitates youth involvement in CBT. However, the specific aspects of the hypothesized relation important to testing the proposed relations are not specified, such as the temporal characteristics of the proposed relation (e.g., whether this relation should be observed within a therapy session or across multiple therapy sessions). In sum, the field would benefit from more refined theoretical models to include a more definitive statement about the expected duration of time between a given therapy process (or set of therapy processes) and the impact on other therapy processes or clinical outcomes.

Hybrid Approach

Therapy process research is increasingly being used to inform dissemination and implementation research (Garland et al., 2006). Indeed, as researchers further develop and refine therapy process measures, they will be able to capitalize on new methods for conducting process research within practice settings. New "hybrid" approaches combine methods from applied services research and the psychotherapy process research tradition (Garland et al., 2006). The "hybrid" approach offers researchers the means to evaluate how EBTs can be delivered within practice settings with integrity and skill. Specifically, this methodology involves a multifaceted assessment strategy that evaluates contextual factors (e.g., organizational climate) that may influence the relation between treatment integrity (adherence, differentiation, competence), relational elements (alliance, involvement), and clinical outcomes. This method provides a framework for studying the dissemination and implementation process. For example, using this "hybrid" approach, researchers found that organizational climate predicted treatment adherence for therapists delivering an EBT for youth disruptive behavior disorders (Schoenwald et al., 2008). This approach provides a roadmap for future efforts to use process research to understand how EBTs can be delivered in practice settings with integrity and skill.

Summary

Therapy process research holds enormous potential for advancing our understanding about how to optimize the delivery of EBTs for emotional and behavioral problems. Clinical trials represent an ideal platform for therapy process research. Future research is clearly needed to address some of the existing theoretical and methodological gaps in the field. However, as the science and measurement of process research progress and advanced statistical methods are increasingly used, process research may play an increasingly important role in future efforts to deliver EBTs in practice settings with integrity and skill.

References

Aarons, G. A. (2005). Measuring provider attitudes toward evidence-based practice: Consideration of organizational context and individual differences. *Child and Adolescent Psychiatric Clinics of North America*, *14*, 255–271.

Barber, J. P., Gallop, R., Crits-Christoph, P., Frank, A., Thase, M. E., Weiss, R. D., et al. (2006). The role of therapist adherence, therapist competence, and alliance in predicting outcome of individual drug counseling: Results from the National Institute on Drug Abuse Collaborative Cocaine Treatment Study. *Psychotherapy Research*, *16*, 229–240.

Barber, J. P., Liese, B., & Abrams, M. (2003). Development of the Cognitive Therapy Adherence and Competence Scale. *Psychotherapy Research*, *13*, 205–221.

Becker, C. B., Zayfert, C., & Anderson, E. (2004). A survey of psychologists' attitudes towards and utilization of exposure therapy for PTSD. *Behaviour Research and Therapy*, *42*, 277–292.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy Theory, Research and Practice*, *26*, 17–25.

Bordin, E. S., Cutler, R. I., Dittmann, A. T., Harway, N. I., Rausch, H. L., & Rigler, D. (1954). Measurement problems in process research on psychotherapy. *Journal of Consulting Psychology*, *18*, 79–82.

Bouchard, S., Mendlowitz, S. L., Coles, M. E., & Franklin, M. (2004). Considerations in the use of exposure with children. *Cognitive and Behavioral Practice*, *11*, 56–65.

Braswell, L., Kendall, P. C., Braith, J., Carey, M. P., & Vye, C. S. (1985). "Involvement" in cognitive-behavioral therapy with children: Process and its relationship to outcome. *Cognitive Therapy and Research*, *9*, 611–630.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Busch, A. M., Uebelacker, L. A., Kalibatseva, Z., & Miller, I. W. (2010). Measuring homework completion in behavioral activation. *Behavior Modification*, *34*, 310–329.

Carroll, K. M., Nich, C., Sifry, R., Nuro, K. F., Frankforter, T. L., Ball, S. A., et al. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug & Alcohol Dependence*, *57*, 225–238.

Castonguay, L. G., Goldfried, M. R., Wiser, S., Raue, P. J., & Hayes, A. M. (1996). Predicting the effect of cognitive therapy for depression: A study of unique and common factors. *Journal of Consulting and Clinical Psychology*, *64*, 497–504.

Chiu, A. W., McLeod, B. D., Har, K. H., & Wood, J. J. (2009). Child–therapist alliance and clinical outcomes in cognitive behavioral therapy for child anxiety disorders. *Journal of Child Psychology and Psychiatry*, *50*, 751–758.

Chu, B. C., Choudhury, M. S., Shortt, A. L., Pincus, D. B., Creed, T. A., & Kendall, P. C. (2004). Alliance, technology,

and outcome in the treatment of anxious youth. *Cognitive and Behavioral Practice*, *11*, 44–55.

Chu, B. C., & Kendall, P. C. (2004). Positive association of child involvement and treatment outcome within a manual-based cognitive-behavioral treatment for children with anxiety. *Journal of Consulting and Clinical Psychology*, *72*, 821–829.

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Coady, N. F. (1991). The association between client and therapist interpersonal processes and outcomes in psychodynamic psychotherapy. *Research on Social Work Practice*, *1*, 122–138.

Cobham, V. E., Dadds, M. R., & Spence, S. H. (1998). The role of parental anxiety in the treatment of childhood anxiety. *Journal of Consulting and Clinical Psychology*, *66*, 893–905.

Crits-Christoph, P., Gibbons, M. B. C., Hamilton, J., Ring-Kurtz, S., & Gallop, R. (2011). The dependability of alliance assessments: The alliance–outcome correlation is larger than you might think. *Journal of Consulting and Clinical Psychology*, *79*, 267–278.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology*, *59*, 20–26.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.

Devellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousands Oaks, CA: Sage Publications.

Donner, A., Birkett, N., & Buck, C. (1981). Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology*, *114*, 906–914.

Doss, B. D. (2004). Changing the way we study change in psychotherapy. *Clinical Psychology: Science and Practice*, *11*, 368–386.

Edelman, R. E., & Chambless, D. L. (1993). Compliance during sessions and homework in exposure-based treatment of agoraphobia. *Behaviour Research and Therapy*, *31*, 767–773.

Edelman, R. E., & Chambless, D. L. (1994). Adherence during sessions and homework in cognitive-behavioral group treatment of social phobia. *Behavioral Research and Therapy*, *33*, 573–577.

Elvins, R., & Green, J. (2008). The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical Psychology Review*, *28*, 1167–1187.

Estrada, A., & Russell, R. (1999). The development of the Child Psychotherapy Process Scales (CPPS). *Psychotherapy Research*, *9*, 154–166.

Feeley, M., DeRubeis, R. J., & Gelfand, L. A. (1999). The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*, 578–582.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., and Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Fjermestad, K. W., McLeod, B. D., Heiervang, E. R., Havik, O. E., Ost, L. G., & Haugland, B. S. M. (2012). Factor structure and validity of the Therapy Process Observational Coding System for Child Psychotherapy—Alliance scale. *Journal of Clinical Child and Adolescent Psychology, 41*, 1–9.

Fleiss, J. R. (1981a). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*, 105–112.

Fleiss, J. R. (1981b). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment*, *7*, 248–260.

Friedlander, M. L., Horvath, A. O., Cabero, A., Escudero, V., Heatherington, L., & Martens, M. P. (2006). System for observing family therapy alliances: a tool for research and practice. *Journal of Counseling Psychology*, *53*, 214–224.

Garland, A. F., Brookman-Frazee, L., Hurlburt, M. S., Accurso, E. C., Zoffness, R. J., Haine-Schlagel, R., et al. (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services*, *61*, 788–795.

Garland, A. F., Hurlburt, M. S., & Hawley, K. M. (2006). Examining psychotherapy processes in a services research context. *Clinical Psychology: Science and Practice*, *13*, 30–46.

Gomes-Schwartz, B. (1978). Effective ingredients in psychotherapy: prediction of outcome from process variables. *Journal of Consulting and Clinical Psychology*, *46*, 1023–1035.

Greenberg, L. S. (1986). Change process research. *Journal of Consulting and Clinical Psychology*, *54*, 4–9.

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, *27*, 637–652.

Hawley, K. M., & Weisz, J. R. (2005). Youth versus parent working alliance in usual clinical care: Distinctive associations with retention, satisfaction, and treatment outcome. *Journal of Clinical Child and Adolescent Psychology*, *34*, 117–128.

Hedeker, D., Gibbons, R. D., & Flay, B. R. (1994). Random-effects regression models for clustered data with an example with smoking prevention research. *Journal of Consulting and Clinical Psychology*, *62*, 757–765.

Henggeler, S. W., & Borduin, C. B. (1992). *Multisystemic Therapy Adherence Scales*. Unpublished instrument. Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, Charleston, SC.

Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology*, *65*, 821–833.

Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance-abusing and -dependent delinquents: Outcomes, treatment fidelity, and transportability. *Mental Health Services Research*, *1*, 171.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (1998). *Multisystemic treatment of antisocial behavior in children and adolescents*. New York: Guilford Press.

Hill, C. E. (1991). Almost everything you ever wanted to know about how to do process research on counseling and psychotherapy but didn't know who to ask. In C. E. Hill & L. J. Schneider (Eds.), *Research in counseling* (pp. 85–118). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hill, C. E., & Lambert, M. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy*

and behavior change (5th ed., pp. 84–136). New York: John Wiley & Sons, Inc.

Hill, C. E., Nutt, E., & Jackson, S. (1994). Trends in psychotherapy process research: Samples, measures, researchers, and classic publications. *Journal of Counseling Psychology*, *41*, 364–377.

Hill, C. E., & O'Brien, K. (1999). *Helping skills; Facilitating exploration, insight, and action*. Washington, DC: American Psychological Association.

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the collaborative study psychotherapy rating scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology*, *60*, 73–79.

Hoagwood, K., Jensen, P. S., Petti, T., & Burns, B. J. (1996). Outcomes of mental health care for children and adolescents: I. A comprehensive conceptual model. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*, 1055–1063.

Hogue, A. (2002). Adherence process research on developmental interventions: Filling in the middle. In A. Higgins-D'Alessandro & K. R. B. Jankowski (Eds.), *New directions for child and adolescent development* (Vol. 98, pp. 67–74). San Francisco: Jossey Bass.

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., et al. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment*, *35*, 137–147.

Hogue, A., Dauber, S., Samuolis, J., & Liddle, H. A. (2006). Treatment techniques and outcomes in multidimensional family therapy for adolescent behavior problems. *Journal of Family Psychology*, *20*, 535–543.

Hogue, A., Dauber, S., Stambaugh, L. F., Cecero, J. J., & Liddle, H. A. (2006). Early therapeutic alliance and treatment outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *74*, 121–129.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *76*, 544–555.

Hogue, A., Liddle, H., Dauber, S., & Samuolis, J. (2004). Linking session focus to treatment outcome in evidence-based treatments for adolescent substance abuse. *Psychotherapy: Theory, Research, Practice, and Training*, *41*, 83–96.

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, and Training*, *33*, 332–345.

Hogue, A., Liddle, H., Rowe, C., Turner, R., Dakof, G. A., & LaPann, K. (1998). Treatment adherence and differentiation in individual versus family therapy for adolescent drug abuse. *Journal of Counseling Psychology*, *45*, 104–114.

Hollon, S., Evans, M., Auerbach, A., DeRubeus, R., Elkin, I., Lowery, A., Kriss, M., Grove, W., Tuason, V., & Piasecki, S. (1988). *Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal psychotherapy, and clinical management pharmacotherapy*. Nashville, TN. Unpublished manuscript.

Hollon, S. D., Munoz, R. F., Barlow, D. H., Beardslee, W. R., Bell, C. C., Bernal, G., et al. (2002). Psychosocial intervention development for the prevention and treatment of depression: Promoting innovation and increasing access. *Biological Psychiatry*, *52*, 610–630.

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., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*, 223–233.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, *38*, 139–149.

Huey, S., Henggeler, S. W., Brondino, M., & Pickrel, S. (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.

Jackson-Gilfort, A., Liddle, H. A., Tejeda, M. J., & Dakof, G. A. (2001). Facilitating engagement of African American male adolescents in family therapy: A cultural theme process study. *Journal of Black Psychology*, *27*, 321–340.

Johnson, S., Hogue, A., Diamond, G., Leckrone, J., & Liddle, H. A. (1998). *Scoring manual for the Adolescent Therapeutic Alliance Scale (ATAS)*. Philadelphia: Temple University Unpublished manuscript.

Jones, E. E., Cumming, J. D., & Horowitz, M. J. (1988). Another look at the nonspecific hypothesis of therapeutic effectiveness. *Journal of Consulting and Clinical Psychology*, *56*, 48–55.

Jones, H. A., Clarke, A. T., & Power, T. J. (2008). Expanding the concept of intervention integrity: A multidimensional model of participant engagement. *In Balance*, *23*, 4–5.

Judd, C. M., & Kenney, D. A. (1981). Process analysis: Estimating meditation in treatment evaluations. *Evaluation Review*, *5*, 602–619.

Karver, M. S., Handelsman, J. B., Fields, S., & Bickman, L. (2005). A theoretical model of common process factors in youth and family therapy. *Mental Health Services Research*, *7*, 35–51.

Karver, M. S., Handelsman, J. B., Fields, S., & Bickman, L. (2006). Meta-analysis of therapeutic relationship variables in youth and family therapy: The evidence for different relationship variables in the child and adolescent treatment outcome literature. *Clinical Psychology Review*, *26*, 50–65.

Karver, M. S., Shirk, S., Handelsman, J. B., Fields, S., Crisp, H., Gudmundsen, G., et al. (2008). Relationship processes in youth psychotherapy: Measuring alliance, alliance-building behaviors, and client involvement. *Journal of Emotional and Behavioral Disorders*, *16*, 15–28.

Kazdin, A. E. (1994). Methodology, design, and evaluation in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 19–71). New York: John Wiley & Sons.

Kazdin, A. E. (1995). Child, parent and family dysfunction as predictors of outcome in cognitive-behavioral treatment of antisocial children. *Behaviour Research and Therapy*, *33*, 271–281.

Kazdin, A. E. (1999). Current (lack of) status of theory in child and adolescent psychotherapy research. *Journal of Clinical Child Psychology*, *28*, 533–543.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*, 1–27.

Kazdin, A. E., & Kendall, P. C. (1998). Current progress and future plans for developing effective treatments: Comments and perspectives. *Journal of Clinical Child Psychology*, *27*, 217–226.

Kendall, P. C., & Beidas, R. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology: Research and Practice*, *38*, 13–20.

Kendall, P. C., Comer, J. S., Marker, C. D., Creed, T. A., Puliafico, A. C., Hughes, A. A., et al. (2009). In-session exposure tasks and therapeutic alliance across the treatment of childhood anxiety disorders. *Journal of Consulting and Clinical Psychology*, *77*, 517–525.

Kendall, P. C., & Ollendick, T. H. (2004). Setting the research and practice agenda for anxiety in children and adolescence: A topic comes of age. *Cognitive and Behavioral Practice*, *11*, 65–74.

Kendall, P. C., & Treadwell, K. R. H. (2007). The role of self-statements as a mediator in treatment for youth with anxiety disorders. *Journal of Consulting and Clinical Psychology*, *75*, 380–389.

Klein, D. N., Schwartz, J. E., Santiago, N. J., Vivian, D., Vocisano, C., Castonguay, L. G., et al. (2003). Therapeutic alliance in depression treatment: Controlling for prior change and patient characteristics. *Journal of Consulting and Clinical Psychology*, *71*, 997–1006.

Klein, M. H., Mathieu-Coughlan, P., & Kiesler, D. J. (1986). The Experiencing Scales. In L. G. A. W. Pinsof (Ed.), *The psychotherapeutic process: A research handbook* (pp. 21–72). New York: Guilford Press.

Kroll, L., & Green, J. (1997). The therapeutic alliance in child inpatient treatment: Developmental and initial validation of a family engagement questionnaire. *Clinical Child Psychology and Psychiatry*, *2*, 431–447.

Liber, J. M., McLeod, B. D., Van Widenfelt, B. M., Goedhart, A. W., van der Leeden, A. J. M., Utens, E. M. W. J., & Treffers, P. D. A. (2010). Examining the relation between the therapeutic alliance, treatment adherence, and outcome of cognitive behavioral therapy for children with anxiety disorders. *Behavior Therapy*, *41*, 172–186.

Luborsky, L. (1976). Helping alliances in psychotherapy. In Claghorn (Ed.), *Successful psychotherapy* (pp. 92–116). New York: BrunnerMazel.

Manne, S., Winkel, G., Zaider, T., Rubin, S., Hernandez, E., & Bergman, C. (2010). Therapy processes and outcomes of psychological interventions for women diagnosed with gynecological cancers: A test of the generic process model of psychology. *Journal of Consulting and Clinical Psychology*, *78*, 236–248.

Marmar, C. R. (1990). Psychotherapy process research: Progress, dilemmas, and future directions. *Journal of Consulting and Clinical Psychology*, *58*, 265–272.

Marmar, C. R., Weiss, D. S., & Gaston, L. (1989). Towards the validation of the California Therapeutic Alliance Rating System. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*, 46–52.

Martin, D. J., Garske, F. P., & Davis, M. K. (2000). Relationship of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*, 438–450.

McLeod, B. D. (2009). Understanding why therapy allegiance is linked to clinical outcomes. *Clinical Psychology: Science and Practice*, *16*, 69–72.

McLeod, B. D. (2011). The relation of the alliance with outcomes in youth psychotherapy: A meta-analysis. *Clinical Psychology Review*, *31*, 603–616.

McLeod, B. D., & Islam, N. Y. (2011). Using treatment integrity methods to study the implementation process. *Clinical Psychology: Science and Practice*, *18*, 36–40.

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*, 541–546.

McLeod, B. D., & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology*, *72*, 235–251.

McLeod, B. D., & Weisz, J. R. (2005). The Therapy Process Observational Coding System-Alliance scale: Measure characteristics and prediction of outcome in usual clinical practice. *Journal of Consulting and Clinical Psychology*, *73*, 323–333.

McLeod, B. D., & Weisz, J. R. (2010). The Therapy Process Observational Coding System for Child Psychotherapy Strategies scale. *Journal of Clinical Child and Adolescent Psychology*, *39*, 436–443.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth*, *4*, 83–86.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340.

National Institutes of Mental Health (1999). *Bridging science and service: A report by the National Advisory Mental Health Council's Clinical Treatment and Services Research Workgroup* (No. NIH Publication No. 99-4353). Rockville, MD: National Institute of Mental Health.

Norton, P. E., Bieler, G. S., Ennett, S. T., & Zarkin, G. A. (1996). Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *Journal of Consulting and Clinical Psychology*, *64*, 919–926.

O'Farrell, M. K., Hill, C. E., & Patton, S. (1986). Comparison of two cases of counseling with the same counselor. *Journal of Counseling and Development*, *65*, 141–145.

Orlinsky, D., Ronnestad, M., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M. J. Lambert (Ed.), *Handbook of psychotherapy and behavior change* (5th ed., pp. 307–389). New York: Wiley.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Piper, W. E., Azim, H. F. A., Joyce, A. S., & McCallum, M. (1991). Transference interpretations, therapeutic alliance, and outcome in short-term individual psychotherapy. *Archives of General Psychiatry*, *48*, 946–953.

Ruma, P. R., Burke, R. V., & Thompson, R. W. (1996). Group parent training: Is it effective for children of all ages? *Behavior Therapy. Behavior Therapy*, *27*, 159–169.

Schoenwald, S. K., Carter, R. E., Chapman, J. E., & Sheidow, A. J. (2008). Therapist adherence and organizational effects on change in youth behavior problems one year after multisystemic therapy. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*, 379–394.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process*, *39*, 83–103.

Schoenwald, S. K., Henggeler, S. W., & Edwards, D. (1998). *MST Supervisor Adherence Measure*. Charleston, SC: MST Institute.

Schoenwald, S. K., Sheidow, A. J., Letourneau, E. J., & Liao, J. G. (2003). Transportability of multisystemic therapy:

Evidence for multi-level influences. *Mental Health Services Research*, *5*, 223–239.

Shirk, S. R., Gudmundsen, G., Kaplinski, H. C., & McMakin, D. L. (2008). Alliance and outcome in cognitive-behavioral therapy for adolescent depression. *Journal of Clinical Child and Adolescent Psychology*, *37*, 631–639.

Shirk, S. R., & Karver, M. (2003). Prediction of treatment outcome from relationship variables in child and adolescent therapy: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *71*, 452–464.

Shirk, S. R., & Saiz, C. C. (1992). Clinical, empirical, and developmental perspectives on the therapeutic relationship in child psychotherapy. *Development and Psychopathology*, *4*, 713–728.

Shrout, P., & Fleiss, J. R. (1979). Intraclass correlations: Uses in assessing interrater reliability. *Psychological Bulletin*, *86*, 420–428.

Silverman, W. K., Pina, A. A., & Viswesvaran, C. (2008). Evidence-based psychosocial treatments for phobic and anxiety disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *37*, 105–130.

Southam-Gerow, M. A., Weisz, J. R., Chu, B. C., McLeod, B. D., Gordis, E. B., & Connor-Smith, J. K. (2010). Does cognitive behavioral therapy for youth anxiety outperform usual care in community clinics? An initial effectiveness test. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1043–1052.

Stiles, W. B., & Shapiro, D. A. (1989). Abuse of the drug metaphor in psychotherapy process-outcome research. *Clinical Psychology Review*, *9*, 521–543.

Stiles, W. B., & Shapiro, D. A. (1994). Disabuse of the drug metaphor: Psychotherapy process-outcome correlations. *Journal of Consulting and Clinical Psychology*, *62*, 942–948.

Teachman, B. A., Marker, C. D., & Clerkin, E. M. (2010). Catastrophic misinterpretations as a predictor of symptom change during treatment for panic disorder. *Journal of Consulting and Clinical Psychology*, *78*, 964–973.

Teachman, B. A., Marker, C. D., & Smith-Janik, S. B. (2008). Automatic associations and panic disorder: Trajectories of change over the course of treatment. *Journal of Consulting and Clinical Psychology*, *76*, 988–1002.

Treadwell, K. R. H., & Kendall, P. C. (1996). Self-talk in youth with anxiety disorders: States of mind, content specificity, and treatment outcome. *Journal of Consulting and Clinical Psychology*, *64*, 941–950.

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology*, *66*, 231–239.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61*, 620–630.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, *5*, 425–433.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*, 200–211.

Weersing, V. R. (2000). *Development and application of the Therapy Process Checklist: Tying process to outcome in child effectiveness research.* Unpublished doctoral dissertation, University of California, Los Angeles.

Weersing, V. R., & Weisz, J. R. (2002a). Mechanisms of action in youth psychotherapy. *Journal of Child Psychology and Psychiatry*, *43*, 3–29.

Weersing, V. R., & Weisz, J. R. (2002b). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology*, *70*, 299–310.

Weersing, V. R., Weisz, J. R., & Donenberg, G. R. (2002). Development of the Therapy Procedures Checklist: A therapist-report measure of technique use in child and adolescent treatment. *Journal of Clinical Child Psychology*, *31*, 168–180.

Weisz, J. R., Southam-Gerow, M. A., Gordis, E. B., Connor-Smith, J. K., Chu, B. C., Langer, D. A., McLeod, B. D., Jensen-Doss, A., Updegraff, A., & Weiss, B (2009). Cognitive behavioral therapy versus usual clinical care for youth depression: An initial test of transportability to community clinics and clinicians. *Journal of Consulting and Clinical Psychology*, *77*, 383–396.

Young, J., & Beck, A. T. (1980). *The development of the Cognitive Therapy Scale*. Center for Cognitive Therapy, Philadelphia, PA. Unpublished manuscript.

Zucker, D. M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement*, *50*, 731–738.

CHAPTER

10 Structural and Functional Brain Imaging in Clinical Psychology

Andrew J. Gerber *and* Marlen Z. Gonzalez

Abstract

The past 20 years has seen a great expansion of research on clinical topics using neuroimaging technology. The work has introduced new theories of mental disorders, supported older concepts, and opened a window into how treatment works on the brain. Neuroimaging is therefore a force shaping clinical psychology in both research and practice. The current chapter introduces issues concerning neuroimaging and clinical psychology, the science and procedures behind the technology, the creation of standards, and future directions.

Key Words: Neuroimaging, clinical psychology, psychopathology, clinical research

Introduction

There are two extreme positions sometimes taken toward neuroimaging in clinical psychology: one of great mistrust, comparing it to phrenology, and one of readiness to believe in anything because it seems so "scientific." Our aim is to offer a middle ground, arrived at by understanding the hard science behind the images and the extent to which the technology can be used to answer questions about psychopathology and treatment.

Readers should take this as an introduction to the current global and specific debates regarding neuroimaging in clinical science. There are a number of theoretical questions still being addressed, and the technology itself is imbued with technical and analytic problems that have yet to be fully disentangled (Malhi & Lagopoulos, 2008; Nemeroff, Kilts, & Berns, 1999; Peterson, 2003). The lack of homogeneity in neuroimaging research standards limits how much can be generalized or synthesized in the literature even within categories of disorders (Bush, Valera, & Seidman, 2005; Etkin, 2010; Sheline, 2003). It may be best to think of this type of research as an exciting frontier, with all of the successes and failures that any new adventure promises.

The chapter has four sections. First we introduce neuroimaging in clinical psychology. Second, we consider the promises and necessary cautions in this line of research, underscoring the limits inherent in the technology. Third, we detail technical and analytic considerations in neuroimaging with specific regard to structural and functional imaging techniques, how they work, their strengths and weaknesses, preprocessing steps, and statistical considerations and implications for clinical trials. Fourth, we highlight current challenges and future directions, discuss standards in clinical neuroimaging, and consider suggestions for best practices.

We hope that readers acquire the necessary tools to become active consumers of neuroimaging studies. It is clear that this technology is changing our models of psychopathology and may someday be a sophisticated tool for diagnosis and treatment management (Malhi & Lagopoulos, 2008; Paulus, 2008). However, we must neither be too eager to look for "confirmation" of what we already believe nor dismissive of the technology's current and

potential contributions. Therefore, even if a clinical researcher never uses neuroimaging, he or she will need to understand the role that neuroimaging plays in the current and new debates of the field.

It is best to become a critical but not reluctant reader of these studies. This critical reader must be armed with a solid understanding of the particular neuroimaging instruments, the methodology and design behind neuroimaging studies, and in particular the technical and statistical considerations necessary for a valid analysis of data. The depths to which such knowledge needs to be understood depends on how involved a researcher may be in the neuroimaging literature.

Promises and Cautions

Correlation, even neural, is not causation. This scientific fact has been drilled into us since our first introductory psychology or statistics course. Yet some of this seems to escape us when we are faced with the crisp monochromatic slices of a living human brain. MacCabe and Castle (2008), in a series of three studies, found that undergraduate participants were more likely to rate an article higher in scientific reasoning if it was accompanied by brain images in comparison to bar graphs or a topographic map of brain activity. Perhaps, as the authors suggest, we judge scientific papers with neural images as somehow more "valid" because it makes concrete an abstract landscape of cognition, not unlike how we value concrete imagery over expository storytelling.

Whatever the reason, the imagery can work against our better scientific training. Each imaging technique uses psychophysical calculations to estimate both structure and functional activity based on particular physiological processes. Said briefly here and covered more in depth later in the chapter: electroencephalograph (EEG) calculates the magnitude of electric signals at the surface of the scalp, positron emission tomography (PET) the intensity of photons emitted by the metabolization of a radioactive tracer, and single photon emission computed tomography (SPECT) the intensity of single photons emitted directly by a radioactive tracer; magnetic resonance imaging (MRI) looks at radio waves emitted by hydrogen molecules; and functional magnetic resonance imaging (fMRI) measures the ratio of oxygenated blood cells to deoxygenated blood cells (Gerber & Peterson, 2008). These are all chemical and physical changes that correlate with the structure or activity in the brain at rest or while performing a task. They offer a valuable but indirect form of observation.

None of these technologies make up for bad scientific practice. As lampooned by Bennet, Baird, Miller, and Wolford (2010), failing to take the necessary statistical and procedural steps leads to faulty data analysis. These researchers designed a study that placed a dead salmon in an fMRI machine and simultaneously ran images portraying different emotions. The "task" was to judge the emotional stimuli. We know the dead salmon could not do anything, but without properly correcting for the multiple comparisons inherent in imaging data (more on this later), the dead salmon appeared to show neural activity associated with individual emotions. If we did not know it was a dead salmon "completing" the task, we might make claims about the neural correlates of making emotional judgment. The specifics of best-practice procedures are contained in the third section of this chapter, "Technical and Analytic Considerations in Neuroimaging."

We introduce these issues here, before diving into the promises and cautions of neuroimaging, to view clearly the limitations of this technology. Knowing these issues sobers us from some of the media hype on neuroimaging and gives us firm ground from which to reach toward its promises.

Neuroimaging in Diagnosis and Assessment

The overarching promise in clinical neuroimaging research is a better understanding of the etiology and progress of psychopathology leading to better, more specialized treatment. Hypothetically this can be reached through understanding sophisticated differences between categories and subcategories of disorders, identifying baseline neurobiological predictors and moderators of therapy outcomes in patients, and establishing neurobiological markers of progress in treatment. We are not near to accomplishing any of these tasks, but some researchers have begun to chip away at the work ahead.

Neuroimaging and Redefining Diagnostic Categories

Although it may seem self-evident now, the idea of psychopathology as a set of conditions with neural correlates was not always appreciated (Fu & McGuire, 1999). Psychological concepts were seen as in "the mind," but not the brain (Roffman & Gerber, 2008). This was justified by supposing that the "hardware" (the brain) was different than its program or "software" (the mind). However, this computer analogy has not held up empirically.

Furthermore, due in large part to the work of the past 20 years in neuroimaging, science has moved closer to models of neural correlates of mental illness. These models have moved from looking at particular structures to looking at neural *systems*, their organization, and the consequences (or lack thereof) in mental illness (Mayberg, 2007). This model does not suggest one particular etiology over another in disrupting these systems. It does suggest that whatever the etiology or etiologies, the result is a cascade of neurobiological differences that accompany a particular mental disorder. In other words, the assumption is that brain differences in structure and function may reflect the emergent consequences of interacting genetic, neurochemical, and environmental components. In this way, neuroimaging can become a tool by which we understand and further differentiate the categories and subcategories of mental illness. The National Institute of Mental Health has already called for a large effort, the Research Domain Criteria (RDoC), to redefine mental illness based on behavior, neural circuits, and genetics. The end result of this research may be the complete restructuring of diagnostic categories based on their neural substrates (Greenberg, 2010). Neuroimaging studies are already weighing in on the redefinitions of disorders for the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorder* (DSM). For example, there are some neuroimaging data that show common and distinct neural substrates in obsessive-compulsive disorder (OCD) compared to other anxiety disorders, thus entering in the debate over whether OCD should be classified as an anxiety disorder (Radua, van den Heuvel, Surguladze, & Mataix-Cols, 2010). Another possibility is the use of neuroimaging to allow diagnosis of bipolar disorder in the absence of a manic episode by understanding the functional and structural differences between major depression and bipolar disorder (Mayberg, 2007). The compilations of neuroimaging research may bring us closer to understanding the mechanisms by which certain disorders tend to be comorbid or express similar symptoms (Kaufman & Charney, 2000; Mayberg, 1994; Plessen, Royal, & Peterson, 2007; Zucker et al., 2007).

One of the challenges in arriving at neural maps of mental illness is that it is difficult to differentiate the neural source of psychopathology from compensatory artifacts. For example, as detailed in Peterson (2003), numerous neuroimaging studies in the 1990s showed "at-rest" hyperfrontality (characterized by higher-than-normal activity in the prefrontal cortex) in OCD. Researchers then surmised that hyperfrontality was the cause of obsessions and compulsions, especially since treatment studies showcased normalization of hypofrontality. However, in a study on voluntary control of tics in Tourette syndrome (TS), Peterson and colleagues (1998) showed that *controlling* tics in TS was correlated with functional hyperfrontality. Subsequently, Peterson and colleagues (2001) gathered structural brain images from children with TS that correlated larger prefrontal cortices in children with *fewer* tics. In other words, these studies suggest that "at-rest" measurements of hyperfrontality in OCD more likely measured the compensatory response of trying to control compulsions rather than a central pathophysiological cause.

In functional neuroimaging there is the risk that neural activity during experimental conditions may reflect uninteresting variability. For example, as detailed in Fu and McGuire (1999), an initial study by Reiman, Fusselman, Fox, and Raichle (1989) showed bilateral activation of the temporal poles in concert with anticipation of an anxiogenic stimuli. However, further data analysis by Drevets and colleagues (1992) localized the activity to muscle contractions at the temples while participants clenched their jaw in anticipation of the unwanted stimuli. As with this last example, methodological improvements may help us untangle central from compensatory neural correlates.

Imaging of high-risk and prodromal populations, conducting research with younger participants, and conducting longer longitudinal follow-ups are all ways to strengthen our ability to make statements on central neural differences across psychological disorders (Bush, Valera, & Seidman, 2005; Marsh, Gerber & Peterson, 2008; Peterson, 2003). However, it is unlikely that there is a one-to-one mapping of brain function or structure to even a simple human experience, let alone a psychological disorder whose effects ripple across dimensions of cognition, emotion, and executive control. While the addition of neurobiological substrates to our understanding of mental illness is an improvement and a boon in terms of legitimizing mental illness as such to the larger public, it is hard to think that a fully neurobiological nosology alone would be more parsimonious and valid than what we have now (Beauregard, 2009). Nevertheless, whether completely moving away from current categorizations of disorders and *restructuring* definitions to neuroscientific explanations or *refining* current categories, neuroimaging is here to affect the nosology of psychological disorders (Malhi & Lagopoulos, 2008).

Identifying Baseline Predictors and Moderators of Prognosis and Treatment Outcome

Implicit in studies on the neural circuits of psychopathology is the idea that such data will help a clinician in deciding what treatments work best for whom—a second promise of neuroimaging. A few preliminary studies have looked at how baseline differences between patients may predict a particular treatment's effectiveness (Table 10.1). Konarski and colleagues (2009) found that depressed responders to cognitive-behavioral therapy (CBT) or a psychopharmacological treatment differed from nonresponders in that they had higher activity in the pregenual and subgenual cingulate cortex at baseline while controlling for symptom severity. Although these findings are preliminary, they are similar to other psychopharmacological studies. For example, in an antidepressant treatment study of 18 hospitalized patients with major depressive disorder (MDD), pretreatment hyperactivity relative to normal controls in the rostral anterior cingulate, as measured by PET, predicted acute treatment response (Mayberg et al., 1997). Nonresponders differed from responders in that they had hypoactivity at baseline relative to normal controls. Another study found anterior cingulate hyperactivity, measured by EEG, as a predictor of antidepressant treatment response 4 to 6 months after initial assessment (Pizzagalli et al., 2001).

In the study by Konarski and colleagues (2009), presumably neural but not symptom differences exposed a baseline predictor of treatment response. However, the small sample (16 per treatment group) limited within-treatment comparisons and imaging sensitivity to small group difference. Furthermore, the study could not coregister (map on) neural activity seen in PET to higher-resolution (more sensitive) structural images. This means that the study could not better check the locations with higher activity (quality control) or assess any volumetric differences. These difficulties are common to clinical neuroimaging studies, limiting the extent that we can generalize findings. Yet, with the increasing level of funding and higher standards in clinical neuroimaging, baseline studies may begin to confidently assess which patients are more likely to improve by conventional interventions. Longitudinal studies scanning the same participants across time may also someday help us obtain a baseline measure of vulnerability to relapse (Mayberg, 2007; Paulus, 2008). However, no imaging modality or technique is currently sensitive enough to capture the individual differences that would recommend one treatment over another.

Projecting the course of mental illness is limited by imaging already symptomatic participants, relying on cross-sectional studies, and our incomplete knowledge of normal brain development (Peterson, 2003). As already mentioned, we cannot disentangle causal from compensatory brain differences in psychopathology through imaging at one

Table 10.1 Neuroimaging Treatment Studies Looking at Baseline Predictors of Treatment Effect

Author, Year	Diagnoses	Treatment	Sample size[1]	Imaging Modality
Forbes et al., 2010	MDD	CBT, CBT + SSRI	13 MDD, 0 NC	fMRI
Ritchey et al., 2010	MDD	CBT	22 MDD, 14 NC	fMRI
Konarski et al., 2009	MDD	CBT, SNRI	24 MDD, 0 NC	PET
Bryant et al., 2008	PTSD	CBT	14 PTSD, 14 NC	fMRI
McClure et al., 2007	AD	CBT, SSRI	12 AD, 0 NC	fMRI
Siegel, Carter, & Thase, 2006	MDD	CBT	14 MDD, 21 NC	fMRI
Buysse et al., 2001[2]	MDD	IPT, IPT+SSRI	46 MDD, 0 NC	EEG

AD, anxiety disorder; CBT, cognitive-behavioral therapy; fMRI, functional magnetic resonance imaging; IPT, interpersonal psychotherapy; MDD, major depressive disorder; NC, normal controls; PET, positron emission tomography; PTSD, posttraumatic stress disorder; SNRI, serotonin-norepinephrine reuptake inhibitor; SSRI, selective serotonin reuptake inhibitor.

[1] Sample size is based on the number of participants included in the neuroimaging analysis and not the original number assigned to groups.

[2] One hundred thirty women were part of a larger treatment study, but only 46 responders, 23 to IPT and 23 to IPT and fluoxetine, were included in the EEG analysis with concurrent sleep measures.

time alone. Cross-sectional imaging studies try to disentangle this question by gathering neural data from a population with a disorder at different ages. This suggests a linear trajectory between neural and symptom changes from childhood to adulthood that may not be so, especially considering that we do not know how many children simply "outgrow" the diagnosis.

Peterson and colleagues (2001), for instance, looked at brain structural images of adults with TS. In direct opposition to the children's brains, adults with TS had *decreased* prefrontal volumes. If we follow a cross-sectional analysis, this would imply that increased prefrontal volumes in childhood (correlated with fewer TS symptoms) would eventually lead to prefrontal volume decreases (more TS symptoms) in adulthood. This is possible, but there are other explanations that a cross-sectional analysis cannot disentangle. The researchers suggested that symptomatic adult participants might have never developed the compensatory prefrontal response to begin with and thus are a different TS subpopulation than the children measured. These adults then might have followed a different developmental course than the children will (Fig. 10.1). Most of the adults who as children did acquire a compensatory response to the tics may have compensated enough to outgrow the TS diagnosis and thus could not participate in the study. Even those who had less concerning symptoms may be less motivated to participate in the study. Of course, this is also a hypothesis that a cross-sectional analysis cannot answer. Our lack of conclusive neurotypical longitudinal data makes it difficult to arrive at well-supported interpretations of such data, although some neurotypical longitudinal studies have been completed (Marsh, Gerber, & Peterson, 2008).

Identifying Biomarkers of Change

Neuroimaging may also one day help us understand the biomarkers of therapeutic change. Some have called for the creation of biological algorithms akin to those used in other medical illnesses to help in the management and decision-making process of mental health treatment (Mayberg, 2003). Among other necessities, this requires a greater understanding of how symptom change and neural change are correlated and how this may be different depending on the therapy used. For example, CBT and an antidepressant may be equally efficacious in treating MDD, but CBT may be correlated with prefrontal cortex changes and antidepressants with limbic system changes. Some clinical neuroimaging studies have observed overlapping but also distinct brain metabolic changes between the use of a structured therapy or a psychopharmaceutical treatment. For example, Buysse and colleagues (2001), Brody and colleagues (2001), and Martin, Martin, Rai, Richardson, and Royall (2001) all looked at the differential effects of interpersonal psychotherapy (IPT) versus using an antidepressant alone or IPT versus IPT and an antidepressant (the first study listed). Martin and colleagues (2001) and Buysse and colleagues (2001) reported general improvement in either condition group (with or without antidepressants) but showed different posttreatment brain activation between the conditions. However, Brody and colleagues (2001) reported greater improvement in the antidepressant group but similar posttreatment functional brain changes.

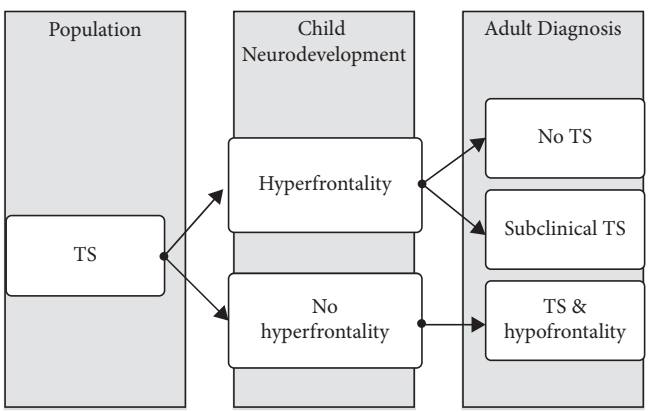


Figure 10.1 Alternate explanation for child hyperfrontality and adult hypofrontality seen in Tourette's Syndrome

There is also reason to believe that the same treatment may manifest different neural changes between categories and subcategories of psychopathology. In a preliminary neuroimaging clinical study, Lehto and colleagues (2008) found that patients with atypical depression (ATD) (n = 8) showed elevated midbrain serotonin transporter (SERT) levels following psychotherapy, but not the typical MDD group (n = 11). However, both groups' symptoms did improve posttreatment as measured by the Hamilton Depression Rating Scale (HAM-D). Therefore, one can imagine, if corroborated by more robust research, that increased SERT could be a marker of therapeutic change for atypically depressed patients, but not for those with typical MDD.

Nevertheless, even if this were true, does the SERT increase really tell us anything more practical than a posttreatment HAM-D can tell us? From a purely clinical point of view, the answer is "no," at this moment. However, correlating symptom changes as rated by standardized scales such as the HAM-D, the Yale-Brown Obsessive-Compulsive scale, and the Abrams and Taylor Scale for Emotional Blunting with neural changes may help us learn what these scales do not tell us: quantitative signs of improvement at the neural level that may not yet have manifested as symptom reduction. Yet a challenge is to differentiate these signs of possible future improvement from neural side effects of going through the treatment.

Incongruent findings seem to more easily display the interpretive challenge inherent in this type of research. For example, a preliminary study looking at cognitive remediation therapy (CRT) and executive function in schizophrenia found frontocortical changes in activity but not corresponding changes in symptom or disability scores (Wykes et al., 2002). Assuming great design and statistical consideration, could this capture a neural precursor to eventual symptom change? Is change in frontocortical activity a moderator—that is, the variable that determines symptom reduction or exacerbation? Or is it a mediator that predicts to what extent CRT may be beneficial (i.e., only those with the lowest baseline frontocortical activity will improve with CRT)? Or is the neural change an artifact of the therapy rather than any process of symptom amelioration? And are neural changes in the absence of eventual reductions in symptoms or disability even clinically meaningful? Since CRT calls for great use of cognitive skills associated with prefrontal areas, functional neural changes captured may only be capturing the use of CRT strategies and not necessarily real-life symptom change. Researchers have sometimes tried to differentiate neural changes related to symptom reduction from changes simply due to use of therapeutic strategies. A functional PET study on the effects of CBT showed posttreatment decreased activity in an area related to sensory processing (thalamus) and increased activity in areas associated with self-regulation (dorsal lateral prefrontal cortex [dlPFC]) in OCD (Saxena et al., 2009). Based on what is known about the neural correlates of OCD, the authors suggested that regulation of sensory input in the thalamus may account for symptom amelioration in any therapy, but that CBT may work through recruitment of self-regulatory neural circuits. These kinds of hypotheses will need to be made and tested to better grasp the extent to which neuroimaging can provide true markers of therapeutic change.

Balancing Promises and Limitations

Balancing the promises of neuroimaging in clinical psychology against the limitations of the technology and our own knowledge is a difficult task. There is no clear road, but as the advent of epigenetics has taught us, there may not be such a thing as a clear road in research. Neuroimaging research in clinical psychology is nascent and promising, a great place to spend time and talent. The next section hopes to give clinical researchers the necessary technical knowledge to participate in the neuroimaging conversation.

Technical and Analytic Considerations

This section will provide an easy reference both to better understand neuroimaging literature and to help in the design of neuroimaging research. Specifically, we consider each modality's uses and limits as they pertain to clinical research. A particular emphasis is placed on structural and functional magnetic resonance imaging because such technology is less invasive than other modalities and has better resolution.

What Is an Image?

Before we understand how individual modalities turn physical data into an image, we must understand the components that define any given neuroimage: pixels, voxels, resolution, and contrast.

A *pixel* is a two-dimensional square that represents a three-dimensional cube of brain tissue called a volume element or *voxel*. This pixel is attributed with a level of grayness raging from black to white that pertains to the numeric calculation of a given physical property used in a particular modality. For

example, in structural MRI a pixel's assigned grayness would be the result of calculating the concentration of hydrogen molecule protons in an anatomically corresponding voxel through radio signals. Pixels are then assembled together (in correspondence with the adjacent voxels they represent) to create a two-dimensional image of a three-dimensional slice of brain tissue. The varying shades of gray in these images then represent the varying levels of a given physical property that is measured by a particular imaging modality. These two-dimensional gray-scale "slices" are what is often shown in a neuroimaging paper.

Neuroimages, like all digital images, are subject to resolution and contrast. *Resolution* is determined by the size and number of pixels in an image. The higher the number of pixels, the smaller the pixel size and the more detailed the image. In neuroimaging a higher number of pixels also means smaller corresponding voxels and greater detail. In the way that lowering the resolution on a digital picture will result in larger pixels and less precision, a lower neuroimaging resolution will mean that a voxel can cover more than one structure. This results in poorer discrimination, as the signals from each structure will be averaged and assigned one corresponding level of grayness.

Related to resolution is the signal-to-noise ratio (SNR). Because a neuroimage is composed of repeated measures of a particular physical property, it is subject to random measurement error or noise within each voxel. A way to decrease noise is to decrease the number of measurements (i.e., decrease the number of voxels and thus increase the size of each voxel). This results in lower resolution and lower image quality. However, a higher resolution will result in more noise and lower signal detection—a more detailed but less accurate image. It is not a small task to determine the best balance between crisp resolution and a higher SNR.

The quality of an image is also determined by its *contrast*. In neuroimaging this is the relative strength of signals coming from each voxel that help discriminate one type of brain tissue from another. Said another way, it is the range of grayness as represented by numbers between black and white in the calculation of the given physical property recorded by a particular imaging modality. The highest contrast is achieved through assigning either pure black or pure white to any given pixel, without any in-between values. As with a black-and-white digital picture, this high level of contrast results in a loss of gradient detail and depth of dimensions. However, too little contrast creates blurred lines and poor differentiation between different structures.

Finally, there is "no free lunch" when it comes to choosing the correct level of resolution, SNR, and contrast. Ultimately, each component will both enhance and degrade an image. What is important is understanding the *purpose* of the image and how manipulation of these components serves that particular purpose. For example, if the purpose of an MR image entails gradient differentiation between levels of myelination (whiteness) in certain cerebral structures versus others, then the ideal contrast will allow for gradual differentiation between highly myelinated and unmyelinated structures.

Structural Neuroimaging

As the name implies, structural neuroimaging generates images of the brain's architecture. Current technologies used for structural neuroimaging are computed tomography (CT), MRI, and diffusion tensor imaging. Other modalities, such as PET, SPECT, and magnetic resonance spectroscopy (MRS), may also be used for structural neuroimaging but are covered in the functional section for their functional capacities. Table 10.2 summarizes these modalities, the physical property they measure, and their strengths and weaknesses.

CT

CT is more often referred to as a CAT scan (computed axial tomography). The instrument consists of an x-ray source that emits a beam through the subject of the scan and an x-ray detector that records x-ray level after it has passed through the tissue. These two devices move around the structure, in our case a living human head sending and recording x-rays from multiple angles per voxel. Like in a conventional x-ray, CT measures the attenuation of the beams at voxels to determine the density of tissue. The result is a black-and-white image showing depth and structural differences. Higher attenuation (denser tissue) results in whiter pixels and lower attenuation results in darker pixels. For example, the cranium will appear very white, while the less dense brain tissue will appear in varying shades of gray. However, CT images have poor spatial resolution, as the range of x-ray absorption is small between tissue types, thus not allowing for great differentiation between structures. Traditionally, CT can image only in axial slices; however, recent helical or spiral CT scanners allow for three-dimensional imaging by stacking two-dimensional slices. In either traditional or spiral CT the attenuation of

Table 10.2 Structural Neuroimaging: Modalities, Measures, Strengths, and Weaknesses

Modality	Measures	Properties Measured	Strengths	Weaknesses
Computed tomography (CT)	Tissue density	Degree of x-ray attenuation in three dimensions as it passes through tissue	Costs, time requirement, spatial resolution	Ionizing radiation, low contrast between soft tissue
Magnetic resonance imaging (MRI)	Chemical environment	Feedback radio signal from water hydrogen nuclei in a magnetic field	No ionizing radiation, high contrast between soft tissue, best spatial resolution	Costs, sensitive to motion, incompatible with ferrous metal (applies to allother MRI technologies)
Diffusion tensor imaging (DTI)	Direction and integrity of neural fiber tracts	Level and direction of water diffusion	Shows direction of a cell and connections between areas of the brain	Spatial resolution, immature image processing and statistical analysis

x-rays enters a set of mathematical equations that result in tomographic reconstruction (of an image). The calculations essentially use attenuation and the location of detectors to create a voxel matrix and geometrically reconstruct a structural image of the brain. Multiple software packages are available with varying CT-reconstruction algorithms. Discussion of the math is beyond the present scope, but it is important to be aware of how physics and math play into this and other imaging modalities.

STRUCTURAL MRI

Structural MRI uses hydrogen atoms found in the water molecules of brain tissue to measure the density of neural structures. The scanner has two magnets: one large and consistently on, and one smaller that turns off and on, emitting variable waves to create a pulse sequence. The bigger magnet serves to align protons in hydrogen molecules toward the magnetic field, and the second magnet disrupts this alignment. In this disruption is contained the quantum-physical information of the shape and makeup of the tissue around these protons. The MRI records these data and uses them to reconstruct an image of the brain through intricate algorithms. This is the basis of all other MRI techniques, including DTI, MRS, and fMRI.

There are advantages to structural MRI: minor invasiveness without radiotracer, excellent spatial resolution, and excellent contrast compared to other modalities. Unlike CT scans, MRI allows for greater differentiation between tissue types and structures. However, the technology is not without flaws. The MRI image is susceptible to movement artifacts and requires both accurate calibrations of head position from the technologist and absolute stillness from the patient during scanning. Also, because an MRI machine has a strong, permanent magnet, no ferromagnetic objects may go in and sometimes not even near the scanner. This includes glasses, coins, ID cards with a magnetic strip, and any metal prosthetics or devices such as older pacemakers. Another issue inherent in MR technology is that images are recreated from radio waves emitted by only a sample and not the entirety of the protons in the tissue. The issue arises because the bigger magnet in the MRI does not align all of the protons in a voxel and some of the protons align against the magnetic field so that when protons "bounce back," they almost cancel each other's signals. Luckily, a few per million molecules are not cancelled out and the MRI collects this information to reconstruct the image (Westbrook & Kaut, 2000). For this reason, MR scanners are said to be less sensitive than other imaging modalities (Peterson et al., 2006). However, MR scanners with higher magnetic fields increase the number of protons calculated in the returning signal, thus increasing image sensitivity, and contrast agents can be used to increase the SNR (Kuo & Herlihy, 2006).

As detailed in Bansal, Gerber, and Peterson (2008), structural MRI can be used to analyze morphological differences between individual brain scans. Morphological MRI requires several preprocessing steps (i.e., before statistical analysis can be performed). Noise can be created by irrelevant radio waves being picked up by the MR machine. Noise can also be created by unplanned fluctuations of the magnets. MR machines will not always run perfectly, and the magnet may interact differently

with the particular participant in the machine. The result is a systematic shift in image intensity causing some pixels that should be white to appear gray, thus blurring some regional borders. However, the frequency of the nonsignal radio waves tend to be higher than those from the protons and can therefore be filtered out from the frequencies used in the image reconstruction. This technique is called "low-pass filtering," and while widely used it is limited, as some anatomic data may also be filtered out (Hou, 2006). However, extension of these methods has preserved some of the efficiency and simplicity of the technique while correcting for lost data or artifacts (Hou, 2006). After using this filter the image will look smoother and more rounded. This is equivalent to using the blur tool on a digital photo and in fact is based on the same mathematical principle, called a Gaussian function. Using the Gaussian blur has several advantages in that it normally distributes the data and effectively reduces the number of multiple comparisons, posing statistical advantages for later parametric testing and correcting for multiple comparisons (Mechelli, Price, Friston, & Ashburner, 2005).

Smoothing the image then allows for the next step, called *segmentation*. Segmentation is the enhancement of differences between various tissue types—bone versus brain tissues, white versus gray matter—so that one can more easily work with the image. This can be done manually through imaging computer programs in the way that one can enhance sections of a digital photo. For example, you can enhance the eye color in a digital photo by selecting the irises using the lasso tool and changing the brightness. In that same way, you can select gray matter (defined as a certain range of grayness per pixel) and enhance it against the white matter it surrounds. There are also a number of software packages that include automated segmentation. In the popular package SPM, segmentation is completed through using statistical probability maps of the spatial distribution of different tissues and identifying tissue types by the voxel intensity distribution (i.e., range of whiteness to blackness) (Mechelli, Price, Friston, & Ashburner, 2005). After segmentation, the tissue is typically "parceled" into anatomically significant regions. For example, subcortical gray matter can be divided into the caudate, putamen, globus pallidus, nucleus accumbens, and thalamus (Bansal, Gerber, & Peterson, 2008). Parceling is like creating distinct layers from a complete image. For example, if you were using clear cellophane sheets, you could place one image of a different part of the brain on each sheet in such a way that when lined up, the sheets would form an entire picture of the human brain. Parceling is doing this process backwards, starting with a full picture and then creating layers that show anatomically relevant subregions from it. This makes it easier to analyze these regions exclusive of those adjacent. Parceling can be done by manually selecting and outlining these regions in each image collected. This is similar to segmentation, but instead of tissue boundaries, structural boundaries are being defined. There are also automated techniques that use statistical modeling based on previous brain images to delineate brain structures. Another common automated function uses landmarks such as the anterior and posterior commissure from which to model the probable placement of other brain structures in relation to the landmarks.

At this point in the preprocessing stage it is difficult to compare individual or group brain scans because of individual and group brain differences. These may be systematic, like in whole-brain size differences based on participant height, or ideographic, such as slightly different orientation of a particular structure in an individual. To make sure that we are indeed comparing the same structures across groups, each brain scan is mapped onto a template brain scan and morphed to fit the standard space. This process, called *spatial normalization*, results in correcting for global brain shape and size differences without morphing the actual individual cortical structures, and thus allowing for comparison (Mechelli, Price, Friston, & Ashburner, 2005). Brains are normalized by rotating images, scaling them to similar sizes, choosing slices that correspond to each other across individuals and groups, and shifting the images across spatial dimensions to line up with each other (Bansal, Gerber, & Peterson, 2008). Automated algorithms for this process are also available that calculate and perform the degree of manipulation across the functions mentioned above (Fischl et al., 2002).

After all four preprocessing steps are completed, two morphological analysis techniques, voxel-based morphometry and analysis of surface contours, are commonly used. The first uses voxel-wise parametric tests to compare local gray matter density between groups and corrects for multiple comparisons (Ashburner & Friston, 2000). However, this test assumes a greater morphological similarity between brains per voxel than is necessarily true, even in neurotypical brains (Fleck et

al., 2008). In contrast, analysis of surface contours assumes that morphological differences in those regions are due to underlying cellular differences at points of interest and focuses on structural shape (Székely, Kelemen, Brechbühler, & Gerig, 1996). Analysis of surface contours looks only at a predetermined region of interest, eliminating whole-brain analysis but allowing for greater sensitivity in smaller structures (Bansal, Gerber, & Peterson, 2008).

DTI

DTI is an MR modality that indirectly measures the direction and integrity of fiber tracks by computing the direction of water diffusion in brain tissue. The underlying theory is that water, in a fibrous structure with a directional orientation, will diffuse more rapidly within the bounds of and in the direction of the fiber tracts than in a perpendicular or contradictory direction (Le Bihan et al., 1986). DTI works by exciting a water molecule with the MR and recording its diffusion from many spatial directions along the directional orientation of myelinated axons (Peterson et al., 2006). Each voxel then contains data both of the speed of diffusion and its direction. When voxel data are combined and computed through reconstructive algorithms, a visualization of the neural fiber tract is created.

DTI does not work in structures without directional orientation as the water molecule is likely to diffuse in any one direction. For similar reasons, DTI has difficulty depicting where neural fiber tracts cross and the boundaries between different types of tissues. Furthermore, DTI has low spatial resolution and is not suited to the study of smaller fiber bundles (Peterson et al., 2006). However, DTI can reliably look at major fiber bundles and has been used to study white matter abnormalities in various psychopathologies (Thomason & Thompson, 2011).

Numerous software packages are available to visualize and analyze DTI data using mathematical models of water diffusivity, the latest being high-angular-resolution diffusion imaging (HARDI) and Q-ball vector analysis (Tuch, 2004; Tuch et al., 2002). However, statistical analysis remains challenging due to lack of standardized processing and analysis procedures, inherent noise and artifacts, and the number of fibers as well as variable physical dimensions in any given analysis. The result is that DTI cannot reliably compare two fibers across individuals or groups and cannot go beyond voxel-wise comparison of more consistent properties (e.g., principal direction) (Peterson et al., 2006).

RELEVANT USES

Structural neuroimaging is frequently used by neurologists in the diagnosis of diseases with prominent anatomic abnormalities. It has, however, also been used in research to find if there are structural brain differences in psychopathology. Some findings include larger orbital frontal regions related to fewer TS symptoms, decreased volume of the hippocampus in posttraumatic stress disorder (PTSD), and decreased volumes in regions of the temporal lobe in schizophrenia (Bremner et al., 1995; Fleck et al., 2008; Peterson et al., 2001). Use of DTI has also shown differences in white matter density in schizophrenia, mood disorder, anxiety disorders, and some developmental disorders (Thomason & Thompson, 2011). A few studies, called perfusion studies, have focused on possible differences in nutrition delivery to different brain tissue due to psychiatric illness (Thebérge, 2008). This research method is adapted from clinical and research work on changes in blood flow and nutrition delivery due to strokes and other similar events. Different from functional neuroimaging, it uses a tracer (either a contrast agent or saturated blood) and measures blood flow as well as blood volume and mean transit time, thus providing a picture of microvascularization (Thebérge, 2008).

ANALYTIC STRENGTHS AND LIMITATIONS

In structural neuroimaging, MRI is generally better than CT because of superior contrast and because it does not expose the participant to ionizing radiation, a fact particularly important when considering neuroimaging in children. It also makes multiple imaging sessions and longitudinal studies more viable. However, MRI technology is more expensive and less readily available. DTI is an exciting new development but best used to examine larger rather than smaller neural fiber tracts. Great caution must be taken in the visualization and analysis of data because the mathematical models are not yet standardized. There is a good amount of research to be done just on figuring out how it works best.

In all of these modalities, finding the best combination of resolution, SNR, and contrast is a balancing act. Chances are that a team, rather than a single researcher, is needed to achieve the best balance.

One important step in volumetric analysis using any of the techniques mentioned is correcting inherent brain size differences in people of different sizes. This is called *morphological scaling* and is part of the "spatial normalization" step in MR preprocessing. Scaling is important because without correcting for body-to-brain size, a researcher may extrapolate inaccurate cognitive correlates from regional brain differences. For example, the famous Paul Broca once concluded that volumetric differences between male and female brains were signs of a better-developed brain and superior intellectual faculties in men, an erroneous finding lampooned by science writer Stephen Jay Gould (Gould, 1978, 1981). Correcting for scaling can be done by linear modeling that takes into account either total body or head size (Arndt, Cohen, Alliger, Swayze, & Andreasen, 1991; Mathalon, Sullivan, Rawles, & Pfefferbaum, 1993).

Another important step, as in data analysis, is correcting for multiple comparisons when reporting brain volumes. Even after scaling to correct for whole-brain size differences, the volume of one particular region of interest is not independent from the volume of another region in the same hemisphere. Furthermore, the size of a particular structure that exists in one hemisphere is not independent from the same structure existing in the other hemisphere. In other words, these volumes are correlated. To account for the intercorrelation of the volumes measured, a repeated-measures analysis of variance is completed (Peterson, 2003). This protects against false-positive errors in detecting hypothesized abnormalities.

Functional Neuroimaging Methods and Specific Data-Processing Guidelines

Functional neuroimaging is the visualization of neural activity in specific brain areas in relation to particular mental functions. One way to scan the neural functions of the brain is simply to have the participant resting without ruminating on one particular topic or another. "Resting-state" methods do not ask participants to perform a particular exercise. Presumably, when compared to neurotypical brain activation, the process can reveal generally dysfunctional neural circuits in those with psychopathology. This is different from paradigms using cognitive tasks, which generally seek to expose a difference in a particular cognitive function rather than whole-brain activity.

In general, one cognitive task alone does not isolate a given cognitive function and its specific neural correlates. This is because multiple processes, from the movement of a mouse to thinking about the task, are happening at the same time, so it is difficult to determine exactly what activity is correlated to what neural activation. To isolate the concept of interest, researchers use what is called the "subtraction paradigm." A *subtraction paradigm* consists of two tasks (control and experimental) that are minimally different from each other, with that difference representing the function of interest. When neural activity differences are observed between the control tasks and the experimental tasks, we assume that it is because of the slight task difference and thus the different cognitive function it requires. This is no easy task, particularly when dealing with higher-order functions. Figure 10.2 shows an example of a subtraction paradigm using the Stroop task.

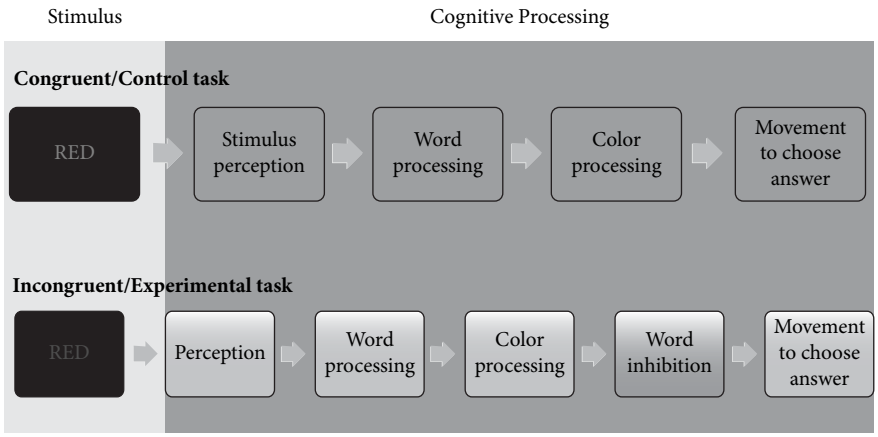


Figure 10.2 Breakdown of subtraction paradigms using the Stroop task.

The current neuroimaging modalities discussed here are EEG, magnetoencephalography (MEG), PET, SPECT, functional magnetic resonance spectroscopy (MRS), and the popular fMRI. Table 10.3 summarizes these techniques and rates their ability to capture where and when neural activity is occurring. Spatial resolution addresses "where" questions and temporal resolution addresses "when" questions. Ideally, we would have high temporal and spatial resolution so that we could determine with great confidence when and exactly where neural impulses occurred, but current technology cannot provide that.

EEG was one of the first neuroimaging modalities. EEG directly measures the electrical activity of cortical neurons through multiple electrodes placed on the scalp. Participants may be asked either to stay relaxed or to perform a task as neural impulses are recorded. In either case it is important to control facial muscles, whose movement may be misinterpreted as the target neural activity. Because EEG is silent, it has a certain advantage over MR technology in recording responses to an auditory stimulus. EEG has excellent temporal resolution, detecting neural change within milliseconds.

Nevertheless, EEG has very poor spatial resolution. More than in MR or PET technology, signals from different sources are blurred together so that it is unclear how many neurons contribute to one set of data. The traditional answer to this challenge has been to add more electrodes. Another technique is to map EEG activity onto structural MR images. The process, called "deblurring," uses mathematical probability models to determine the area correlated to the EEG activity observed (Gevins, Le, Leong, McEvoy, & Smith, 1999).

Table 10.3 Functional Neuroimaging: Modalities, Measures, Strengths, and Weaknesses

Modality	Measure	Properties Measured	Strengths	Weaknesses
Electroencephalography (EEG)	Neural activity in the cerebral cortex	Electrical signals at scalp surface	Costs, best temporal resolution, no ionizing radiation, portable equipment	Spatial resolution, no subcortical measurement
Magnetoencephalography (MEG)	Intraneuronal current flow in cortical cells	Magnetic fields of action potentials atscalp surface	Temporal resolution, no ionizing radiation	Costs, spatial resolution (better than EEG), usually no subcortical measurement
Positron emission tomography (PET)	Cerebral blood flow energy consumption, or neurotransmitter system components	Photons emitted from collision between decaying radiotracer's positrons and body's electrons	Customizable radiotracers allow for measurement of specific CNS activity	Costs, ionizing radiation, spatial and temporal resolution, requires injection of foreign substance into participant
Single-photon emission computed tomography (SPECT)	Concentration of a specific tracer/tracer consumption	Intensity of single photons emitted from decay of radiotracer	Costs (compared to PET), customizable radiotracers	Ionizing radiation, lower spatial resolution, lower temporal resolution than PET
Magnetic resonance spectroscopy (MRS)	Concentration of specific brain metabolites	Feedback radio signals and spectral peaks at different radiofrequencies	Costs (relative to PET), no ionizing radiation, measures brain metabolites	Spatial and temporal resolution, time requirement, limited measurable metabolites, relative concentrations only
Functional magnetic resonance imaging (fMRI)	Neural activity based on oxygen blood ratio	Shifts in blood-oxygen-dependent response (BOLD)	Better spatial resolution than EEG, no ionizing radiation, no injection	Low temporal resolution, usually measures relative activity, difficult paradigm development and analysis

EEG has a theoretical weakness in that it supposes a single source of brain activity. This single source hypothetically "sets off" any subsequent neural activity across the cortical surface. However, it is mathematically impossible to find the theoretical source (Gevins, Le, Leong, McEvoy, & Smith, 1999). Moreover, other functional imaging has shown that the brain can have simultaneous activity at different locations without necessarily originating from a single source. Finally, EEG can measure only the cortical surface and not subcortical activity. Much of the current conversation in psychopathology centers around cortical–subcortical connections, and EEG research is limited in how much it can say about those connections (Peterson, 2003).

As with structural imaging, functional imaging involves preprocessing steps that try to increase SNR. Processing steps of EEG data attempt to correct for noise created by unintended muscle movement by the participant or random EEG spikes that can be caused by atmospheric changes. Some of these errors have known properties that can be filtered out through algorithms that separate and remix the data. An automated form of these statistical processes, FASTER (Fully Automated Statistical Thresholding for EEG Artifact Rejection), was developed by Nolan, Whelan, and Reilly (2010).

MEG is similar to EEG, but it measures the magnetic fields of action potentials rather than direct electrical currents. Due to provisions of that measurement, the result is better spatial resolution, with comparable temporal resolution, and less vulnerability to motion artifacts (Peterson et al., 2006). However, unlike EEG, MEG necessitates a magnetically shielded room and expensive, bulky hardware. Furthermore, MEG suffers from the same theoretical constraint that a single source originates brain activity associated with a given task.

MEG preprocessing steps involve filtering noise from unintended muscle movement by the participant (usually eye movement) and cardiovascular functions. These are filtered through algorithms that remove or dampen certain signal ranges associated with this common noise. As with EEG, MEG spatial resolution can also be enhanced or "deblurred" through the use of MRI (Peterson et al., 2006). However, this also greatly increases the costs and logistical considerations in using MEG.

PET works by recording the signal from the interaction between an unstable nuclear isotope called a radiotracer and other particles in the brain. For example, one common radiotracer is fludeoxyglucose (FDG), a radioactive molecule that is much like glucose and is thus treated similarly by the brain. FDG is injected in a participant through the vein. The radiotracer follows the path of glucose and decays at a known rate, emitting a positron that collides with an electron in surrounding tissue and generates two photons (gamma rays) that travel in opposite directions out of the subject's body. Detectors positioned around the participant pick up the intensity and source of these gamma rays. More highly activated neurons need more glucose, so the FDG theoretically reveals which areas are more active than other. If measured "at rest," we can compare differential baseline activity between participants. In a cognitive task paradigm, we can match differential neural activity with particular cognitive actions and make inferences about what areas of the brain support those actions.

PET has several advantages in that it can measure both "at-rest" brain activation and activation during cognitive tasks. Moreover, the radiotracers used are specific and sensitive since they are analogues to chemicals in the brain. However, these radiotracers have short half-lives, necessitating an on-site cyclotron (particle accelerator that creates them). Cyclotrons are expensive to maintain, thus increasing the costs and time spent on PET scanners. Furthermore, the use of ionizing radiation is inadvisable with certain populations, such as children, and limits the number of scans that can be performed on one single individual (Peterson et al., 2006).

PET requires a number of corrections due to its vulnerability to movement artifacts, detector cooldown or differences in sensitivity, and photons scattering rather than traveling at the expected 180-degree angle. Like CT, PET images are also reconstructed through algorithms that fall under tomographic reconstruction. Reconstruction methods include algebraic reconstruction (ART), Lanweber, conjugated gradient, and expectation-maximization maximum-likelihood (EM-ML) (Gregor & Huff, 1997). PET imaging can be combined with MR or CT structural data to increase the spatial resolution of PET and increase the SNR (Burger et al., 2002; Woods, Mazziotta, & Cherry, 1993).

SPECT is similar to PET in that it also measures gamma rays. However, unlike PET radiotracers, SPECT radiotracers emit gamma rays that are directly measured as they decay rather than through measurement of positron annihilation. This results in poorer spatial resolution in SPECT than in PET. Nevertheless, SPECT is less expensive than PET because SPECT radiotracers have longer half-lives and thus do not necessitate an on-site cyclotron.

MRS uses the same principles of MRI to record biochemical signals in the brain. In its simplest form MRS does not actually produce an image, but rather a frequency that shows the concentration and distribution of metabolites relative to a reference signal. The reference signal can be water in brain tissue, an external water vial by the participant's head, or another metabolite such as creatine (Peterson et al., 2006). A frequency is marked by peaks called chemical shifts that correspond to the position of the metabolites relative to the reference. The area below a peak is a marker of intensity and a quantifier of the relative concentration of the corresponding metabolite. These measurements can then be mapped onto structural brain scans to create a two-dimensional or three-dimensional voxel matrix of metabolite concentration in the brain.

The most common metabolites measured in MRS are N-acetyl-aspartate (NAA), total creatine (tCr), and choline-containing compounds (tCh). These are easier to measure than other metabolites because they exist in a higher concentration and have discrete frequency peaks. Some studies have looked at glutamate or GABA in psychopathology (Chang, Cloak, & Ernst, 2003; Courvoisie, Hooper, Fine, Kwock, & Castillo, 2004; Moore et al., 2006). However, these metabolites are harder to measure because of their low concentration, weak signals, and complex structures (Gerber & Peterson, 2006).

Limitations to MRS and MRSI include lower spatial and temporal resolution compared to other MRI technologies, low SNR, and use of relative rather than absolute quantification of metabolites. MRS is best suited to higher magnetic fields (above 1.5 Tesla) because delineations between chemical shifts are dependent on the strength of the field (Maier, 1995).

fMRI uses the same technology as structural MRI to collect information regarding the level of oxygen content in brain tissue at systematically variable times. The biological physics behind this is that greater consumption of glucose by active neurons results in a change in the ratio between oxyhemoglobin and deoxyhemoglobin. When neurons activate, blood capillaries open and release more oxygenated blood, thus decreasing the relative presence of deoxygenated blood around those neurons. Deoxyhemoglobin creates magnetic field distortions and the higher presence of oxygenated blood lessens this effect, slightly increasing the magnetic resonance signal (Buxton, Uludag, Dubowitz, & Liu, 2004). Because in fMRI the signal is dependant on higher relative levels of oxygen, it is termed a *blood-oxygen-level-dependent (BOLD) response* (Ogawa et al., 1992). The BOLD response provides indirect measurements of neuronal activity, but not activity itself as with EEG (Fig. 10.3).

fMRI is possibly the most popular of the functional neuroimaging modalities because of its low invasiveness (no radiotracer), fair ability to study cortical and subcortical areas, and excellent spatial resolution. It also has a comparatively intermediate temporal resolution that can be manipulated to suit the goals of a particular study (Huettel, Song, & McCarthy, 2004). Subtraction paradigms used in fMRI allow for targeted exploration of a particular mental function but are based on assumptions that have been challenged before neuroscience (Peterson et al., 2006)—namely that cognitive functions and brain activity are additive; as cognitive functions increase in complexity, brain activity theoretically increases in a hierarchical and linear fashion. Neither of these assumptions is unchallenged, and they seem unlikely since brain processes themselves do not appear to be linear (Nemeroff, Kilts, & Berns, 1999). More of this is explained in consideration of all functional imaging techniques below.

Other weaknesses to fMRI include multiple challenges to a high SNR. BOLD signal amplitude is not robust and decreases as higher cognitive functions are introduced (Huettel, Song, & McCarthy, 2004). Furthermore, noise is created by thermal fluctuation, head motion, autonomic nervous system changes, and unrelated neural activity. In summary, a good fMRI study achieves the right combination

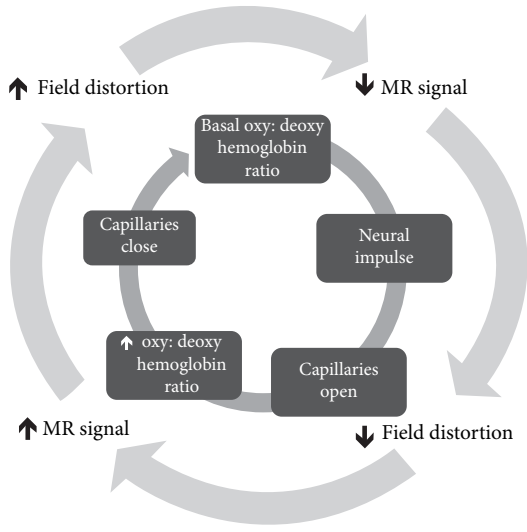


Figure 10.3 Blood-oxygen-level-dependent (BOLD) response and fMRI signal.

of a well-thought-out subtraction paradigm that exploits small significant changes between conditions, good scanning practices, proper data preprocessing steps, and appropriate statistical corrections. All of these are necessary.

fMRI preprocessing steps attempt to increase the SNR by correcting for errors, normalizing data, and filtering unrelated variability in preparation for statistical analysis (Strother, 2006). For example, brain slices covering the entirety of the brain are not all scanned at the same time, meaning that slices next to each other do not represent the same state of the brain at exactly the same time. To correct for this a technique called *temporal interpolation* is used. It uses mathematical modeling and information from nearby points in time to estimate signal amplitude at the time in question. Head motion is corrected through realignment algorithms and spatial interpolation, which uses spatial data from nearby locations to estimate the signal amplitude had there been no head motion (Huettel, Song, & McCarthy, 2004). fMRI data are also mathematically mapped to higher-resolution structural images to increase spatial resolution (Peterson et al., 2006). As in structural MRI, these images are smoothed, averaging functional signals of nearby voxels, and normalized to a similar brain atlas. This corrects for some of the noise and irrelevant differences between participants' brains respectively. Finally, known noise signals such as breathing or eye movement are filtered out of the data (Huettel, Song, & McCarthy, 2004). Several preprocessing software packages are available (Table 10.4), some of which work during scanning (Cox, 1996). Although there are some "common" preprocessing steps, their order and importance are not standardized (Strother, 2006).

After preprocessing steps are completed the data are ready for statistical analysis. Comparisons are generally completed on a voxel-based, or region of interest (ROI)-based, analysis (Peterson, 2003). As the name implies, *voxel-based analysis* compares signal changes across groups, one corresponding voxel at a time. This inherently makes voxel-based comparison particularly susceptible to multiple comparison errors. On the other hand, *ROI-based analysis* involves defining boundaries to anatomic structures of interest and blurring the signal changes by taking the average per-voxel change and either summing or averaging for the entire ROI, leaving one data point for this region (Huettel, Song, & McCarthy, 2004). Voxel-based analysis is best suited for whole-brain exploration, whereas ROI-based analysis is preferred when a particular brain region is the subject of the hypothesis. The weaknesses of these analysis are in

Table 10.4 Common Preprocessing Steps in Functional Magnetic Resonance Imaging

Steps (in common order)	Short Description
Motion correction	Using coregistration algorithms, realigns brain slices to a common reference
Slice-timing correction	Using temporal interpolation, models brain activity as if all slices were captured at the same time
Temporal filtering	Using various techniques such as high-pass or low-pass filtering, corrects for irrelevant MR drift across time and physiological noise
Spatial filtering/smoothing	Using a Gaussian kernel, assigns new pixel value through weighing of neighborhood pixel values, decreasing noise and false-positive errors
Geometric unwarping	Using "field mapping," corrects for inhomogeneity in the magnetic field of the scanner
Spatial normalization	Using a brain atlas to register participants' brain activity, corrects for irrelevant anatomic differences and creates a common anatomic reference. Begins with coregistration of a participant's functional data to his or her structural scan.

List of common fMRI preprocessing and analysis software

Analysis of Functional Images (AFNI): http://afni.nimh.nih.gov/afni

Brain Voyager (BV): http://www.brainvoyager.com

Software Library: http://www.fmrib.ox.ac.uk/fsl

Parametric Mapping: http://www.fil.ion.ucl.ac.uk/spm

their underlying assumptions: voxel-based analysis assumes that after preprocessing the structure and function of each brain is the same across individuals, and ROI-based analysis assumes that we can reliably define regions. The first assumption seems unlikely, especially when we consider comparing neurotypical versus neuropsychiatric brains (Marsh, Gerber, & Peterson, 2008). The second assumption is equally shaky because, as previously discussed, it is difficult to precisely define anatomically and functionally relevant regions using MR images (Peterson, 2003).

Finally, fMRI is particularly vulnerable to error due to multiple comparisons, although as discussed multiple comparisons do present a statistical challenge to some extent across all modalities. In the statistical analysis of fMRI data, multiple independent tests are conducted per voxel, increasing the likelihood of false-positive or type I errors. The solutions are corrections like the Bonferroni correction, which decreases the alpha value proportionately to the number of independent statistical tests conducted (Huettel, Song, & McCarthy, 2004). However, the Bonferroni correction may not be as appropriate for whole-brain analysis because it may require a p value so small ($<1 \times 10–^{5}$) that type II errors are likely to occur. Other less conservative tests take into consideration the spatial relationship of possible findings. For example, if significant voxels are all clustered around a single point, it is more likely that they represent a "real" finding than if they are scattered randomly across the brain. The false discovery rate (FDR) is a commonly used statistical technique that corrects for multiple comparisons by estimating the likelihood of type I errors (Genovese, Lazar & Nichols, 2002). Another commonly used technique is the finite impulse response (FIR) model, which averages BOLD signal at different time intervals, effectively reducing the number of individual data points (Lindquist & Wager, 2007).

RELEVANT USE

Functional neuroimaging is an exciting tool for clinical research as it gives us a window through which to see the neural makeup of disorders and a new way to evaluate treatment interventions (see Chapter 1 in this volume). Concepts that were previously difficult to study, such as emotional regulation, implicit learning, and social cognition, have all benefited from the use of functional imaging (Roffman & Gerber, 2008). As we understand more about the neural correlates and networks supporting normal processing, we also understand abnormal processing and possible vulnerabilities as well as targets for therapy. Functional neuroimaging, particularly MR technology, has been instrumental in probing the regulatory deficits in disorders from mood and anxiety disorders to schizophrenia (Drevets, 2001; Etkin, 2010; Wolkin et al., 1992). Well-designed and well-implemented studies can inform theories about mental disorder and lead the field closer to etiologies. They are also the key to developing clinical applications that will allow for both more individualized and more empirically grounded care.

ANALYTIC STRENGTHS AND LIMITATIONS

The great analytic strength in functional neuroimaging modalities is that they give access to brain differences that may not be apparent on structural imaging alone. Furthermore, in combination with structural neuroimaging, functional imaging is a window through which we may understand neuroplasticity. We know now that even "mundane" activity such as playing the piano can change the brain (Bengtsson et al., 2005). Through longitudinal imaging studies we may one day understand how functional differences at an early age can lead to structural differences later in life.

In general, all functional imaging techniques have poorer spatial resolution than the detailed MR images, and the ability to register activity to structural images has greatly improved the strength of functional analyses. Each modality uses different physiological activities to infer when and where neural activation occurs. This makes it difficult to cross-analyze data. However, researchers have begun to use these differences to capitalize on the strengths and limit the weaknesses of individual modalities by combining them in a single imaging session. For example, fMRI has poor temporal resolution and good spatial resolution, while EEG has excellent temporal resolution but poor spatial resolution. There has been some interest in combining these technologies to counteract the individual shortcomings (Goldman, Stern, Engel, & Cohen, 2000). Nevertheless, EEG-MRI and other multimodal functional imaging types are even more nascent than the imaging field itself and require theoretical and computational advances (Ritter & Villringer, 2006).

A new development in functional imaging has been the attempt to map the distinct neural *networks* that support different cognitive functions. Called "functional connectivity," this analytic technique uses structural and functional MR data to investigate links between relevant structures (Rykhlevskaia, Gratton, & Fabiani, 2008). While

the functional data determine the synchronization of different activity, structural data narrow down which of these activities are actually related to the same cognitive function. The assumption is that functional networks are more likely to be structurally connected in some way. The technique is fairly new but presents with the possibility of establishing "at-rest" connectivity through MR technology. Recently, some researchers have even used it to predict individual functional brain development (Dosenbach et al., 2010).

The usefulness of functional imaging in research is limited by the subtraction methodology used. As mentioned briefly, there are two assumptions built into these paradigms, the first being that brain activity is additive and follows a linear pattern. In other words, the harder the task, the more brain activity in the neural networks supporting that task. However, it is actually unclear what more or less brain activity actually means. One can interpret less brain activity in a particular area as a failure to recruit appropriate resources or as a more efficient use of resources. Poor performance and excellent performance on a task can both be correlated with less activity than moderate performance on a task (Jackson, Briellmann, Waites, Pell, & Abbott, 2006). The second assumption is that the addition of an extra task to the baseline cognitive task also adds an extra cognitive process that does not interact with the previous process used. This "pure insertion" hypothesis has also been critiqued and modified successfully (Sternberg, 1969).

Functional imaging places great emphasis on the appropriate matching of control versus experimental conditions (Peterson, 2003). This is easier when looking at more basic cognitive functions such the difference in neural activity between seeing letters or recognizing a word. But it becomes much more difficult when we try to study higher-level cognitive processes such as emotions or decision making. For example, a paradigm asks a subject to determine the appropriateness of given solutions to both impersonal or personal moral dilemmas and nonmoral dilemmas (Greene, 2001). The control or baseline condition is the nonmoral dilemma, representing the baseline cognitive processes and brain activation involved in considering the appropriateness of a certain action in a difficult situation. To that is added the "extra" moral component, which comes in both an impersonal and a personal variety. Ideally, neural differences between moral and nonmoral dilemmas reflect the added process of the moral component, and neural differences between personal and impersonal conditions reflect differential processing of the moral dilemma based on emotional closeness. This premise is hard to execute because even the quality of the writing in each of these conditions can influence neural differences. Differential brain activity can be due to different emotive language across conditions, use of known or unknown characters, and cognitive processing requirements across conditions (McGuire, Langdon, Coltheart, & Mackenzie, 2009).

Functional neuroimaging also has wide intrasubject and intersubject variability that increases with task complexity. This is because as tasks become more complex, there is greater flexibility (inside or outside of the participant's awareness) in the strategies used to complete a task. This is apparent in the well-known speed-versus-accuracy tradeoff, where a given participant may choose accuracy over speed or vice versa and even change the strategy at different points in the scan (Huettel, Song, & McCarthy, 2004).

Considerations for All Clinical Neuroimaging Research

There is a lot to consider when designing a neuroimaging study, and even more to consider when dealing with the myriad of variables inherent in most mental illnesses. The theoretical foundations of the imaging technique itself, sample population, task used, preprocessing steps, and statistical functions all interact and ultimately affect the analysis. Much is yet to be determined in clinical neuroimaging research, but there are some things we know that can move us forward in terms of technical design: better sample construction, using *a priori* theory-driven hypotheses, and differentiating mediators from moderators.

As in all clinical research, recruiting a representative sample presents challenges. Inpatients are easier to reach, but these technologies are susceptible to movement artifacts, so participation is sometimes limited to those who can tolerate the procedures. Clinical research alone is also expensive, more so when one adds a neuroimaging component, particularly PET or MRI. This makes it difficult to recruit the right participants and the right number of participants. Nevertheless, without controlling for *who* is studied and *how many*, it is difficult for clinical neuroimaging studies to go beyond preliminary implications. Susceptibility to multiple-comparison errors is best addressed through a combination of statistical corrections and larger sample size. Neuroimaging treatment studies should aim to follow randomized

controlled trials (RCTs; see Chapter 4 in this volume) in terms of having a control group, an experimental therapy group, and a waitlist group. A good marker is to have at least 30 participants per condition. As in RCTs, it is important to have well-thought-out inclusion guidelines and to accurately document comorbidity as well as any kind of drug use, since these may affect the analysis.

Although clinical neuroimaging is new, it is grounded in a fairly robust literature of neuroscience. Furthermore, we do have some leads on what areas could be implicated in particular disorders (see Etkin, 2010; Frewen & Lanius, 2006; Kaye, 2008; Marsh, Gerber, & Peterson, 2008; McCloskey, Phan, & Coccaro, 2005; Rapoport & Gogtay, 2010; Sheline, 2003, for review). In other words, we are at a stage where we can make some precise hypotheses on what neural change we can expect after treatment in a particular disorder. There is certainly a place for wider exploration, but if the goal is to find reliable biological markers of illness and remission, some narrowing down must occur. *A priori* hypotheses give more weight to final analyses, leaving studies less susceptible to false-positive errors (both mathematically and because we are forced to narrow our sights), and speak more to the theory behind the observable change. As in all research, it is not enough to see a change and then postulate how it happened.

Finally, we offer a rather conceptual point, but a point that does affect the technical aspects of neuroimaging in clinical research. There are different ways in which variables may account for the differences in people's behavior. Third variables can act as either (a) a moderator that affects under what circumstances or to what extent a treatment may influence the outcome measure or (b) a mediator, a mechanism through which the independent variable affects the dependent variable (Baron & Kenny, 1986; see also Chapter 15 in this volume for consideration of mediators and moderators). In clinical neuroimaging research, neural correlates are third variables that may be either characteristics that determine the extent to which treatment alleviates symptoms or mechanisms of change central to the disorder. The field has not really discussed when a neural variable should be treated as either a mediator or a moderator, but examination of the relationships among independent and dependent variables and third variables should be consistent with existing terminology and analytic strategies detailed broadly for RCTs (Kraemer, Wilson, Fairburn, & Agras, 2002).

Current Challenges and Future Directions

Clinical neuroimaging research stands to provide great advancement in the understanding of the pathophysiology underlying psychological disorders. It may one day lead to better diagnostic criteria, better differential therapeutics, and biological markers of treatment response. However, imaging studies of clinical interventions is nascent and ripe with as-yet-unresolved issues. These issues are multidisciplinary across physics, physiology, mathematics, and psychology. No one person or field may be able to solve these. Nevertheless, clinical neuroimaging research can do its part by creating and maintaining study standards.

The understanding of RCT design is essential to creating these standards (see Chapter 4 in this volume). The American Psychiatric Association's Committee on Research on Psychiatric Treatments created a set of guidelines to assess quality in therapy-specific RCTs (Kocsis et al., 2010). These guidelines include ratings for the description of participants, definition of delivery of treatment, outcome measures, data analysis, treatment assignment, and overall quality of the study. All of these are necessary to control for confounding variables as well as to counter some of the weakness inherent in neuroimaging.

For example, under the description of participant criteria, among two other factors, the checklist calls for description and justification of inclusion and exclusion criteria and description of relevant comorbidities. This is paramount in neuroimaging research as we do not yet fully understand how symptom severity or comorbidity is related to brain structure and function. Undesired heterogeneity, perhaps more than in nonimaging trials, sorely undercuts the inferences we can make from the data on a particular disorder.

Other design quality considerations that are important to all RCTs, but that must be particularly emphasized in clinical neuroimaging, are adequate sample size, control groups (including waitlist group), appropriate statistical tests, and blinding. The "gold standard" for RCTs involves a treatment group, a normal control group, and a waitlist group, with at least 30 participants per group. Given the immense costs of RCTs and neuroimaging, these standards may not seem feasible. However, there is a case to be made regarding the potential returns in terms of knowledge, prevention, treatment enhancement, and possible investment returns (Johnston, Rootenberg, Katrak, Smith, & Elkins, 2006). For example, it

was largely through the work of neuroimaging that TMS treatment for chronic depression was created and modified to target specific brain structures (Mayberg et al., 2005). Moreover, continued accumulation of underpowered neuroimaging studies will be of little use for the field, given the impact of power on the stability, generalizability, and interpretability of estimates (see Chapter 12 in this volume for a full discussion of statistical power in clinical research). Correct use and proper reporting of statistical analysis tools may seem more obvious in light of their importance in neuroimaging, but this is not always done (Bennett, Balnd, Miller, & Wolford, 2010).

Finally, blind independent raters of outcome measures have not always been used or are not described in neuroimaging treatment studies. For example, one study on social phobia had the same therapist rating the outcome measure and deciding at what point the patient was thought to be a treatment responder (Furmark et al., 2002). This is a problem in any RCT, but in a neuroimaging RCT, false positives can be common and lack of control at the front end worsens this problem.

Guidelines for standards in therapy-related RCTs are a good start to better standards for clinical neuroimaging studies, but not the end. Although a lengthy discussion of imaging-specific clinical research standards cannot be contained here, we propose a few practices to increase the quality of neuroimaging treatment studies.

Documentation of Proper Data-Processing Functions Executed

Preprocessing steps are important and sometimes crucial to understanding results and analysis. For example, in a review of structural neuroimaging of mood disorders, the author noted that, of the researchers who had looked at hippocampal volumes in major depression, those who had negative findings used lower resolutions than those who had found differences (Sheline, 2003). Included in this documentation of preprocessing steps are demarcation of structural boundaries, sampling of neural slices, and brain normalization procedures used.

Thorough Description of Novel Functional Imaging Paradigm and its Basis or Use of Established Paradigm

With advances in our understanding of neuroimaging technology's limits, paradigms should now be more highly scrutinized. When not using a previously tested paradigm, it is important to thoroughly describe the conditions and the rationale behind them. Peterson (2003) suggests that creating multiple subtraction paradigms that test the processing behind one mental act from the simplest to the most complex can help disentangle compensatory from central neural abnormalities. They will also test the assumption of cognitive serial processing, the bedrock of subtraction paradigms. This is possible only through an open and thorough explanation of paradigms that lead to scientific criticism and progress.

Clear Statement of A Priori Neural Areas of Interest and Defined Hypotheses; Clear Differentiation with Unanticipated Findings

Any one neural structure or neural circuit may be activated in multiple and seemingly unrelated functions. It is therefore easy to see a statistically significant structural or functional difference and then superimpose what this means ad hoc based on what is "known" about those significant locations. Instead, researchers should arrive, as with any other research, with a theoretical and physiological basis for expected results. If an unexpected difference has been found in an area not considered in the original hypotheses, this must be differentiated in the study's reporting. This ensures that we are not looking for just any difference, but a particular theoretically significant difference (or sameness) that feeds theory, which feeds better imaging designs.

Blind Neuroimaging Data Analysis

We know much about experimenter bias, yet when it comes to clinical neuroimaging studies, it is unclear who analyzes the data and whether they know the anticipated outcomes or not. We suggest a standard of using an independent researcher to de-identify the groups and another blind researcher to reconstruct and analyze the neuroimaging data within *a priori* protocols. For example, one independent researcher will mask the treatment, control, and waitlist groups using letters and then hand it off to another researcher. This researcher will then carry out protocols for analysis (appropriate preprocessing and data filtering, ROI or voxel-wise comparison). Another independent researcher/clinician would evaluate participants as per RCT guidelines (Kocsis et al., 2010). Blinding is often logistically hard in RCTs, but this does not have to be the case for any neuroimaging portion of treatment studies.

Development of Standardized Paradigms and Screening for Looking at Process of Change in Therapy

Instead of continuously developing new paradigms to probe the same functions, we may better use our energy in modifying and studying the validity of available probes. This begins with developing proper stimuli and tasks. For example, building from the work of Erwin and colleagues (1992), Gur and colleagues (2002) created a facial emotion processing task with stimuli that displayed racially diverse faces displaying a range of common expressions (sad, happy, fearful, etc.). The images were tested with normal controls to ensure that emotions could be properly identified (Gur et al., 2002). These stimuli have been used in schizophrenia, anxiety, and mood disorder neuroimaging research (e.g., Bryant et al., 2008; Kohler et al., 2003; Pavuluri, Oconnor, Harral, & Sweeney, 2007). Bush, Valera, and Seidman (2005) suggest that the use of identical paradigms and parameters across psychopathologies will facilitate direct comparisons among and between disorders.

Conclusion

The refinement of clinical neuroimaging research is likely to lead to both novel and more refined treatment. In addition, the advances in clinical neuroimaging may influence the use and management of concurrent psychopharmacological and therapeutic treatments, perhaps leading to more effective use of therapeutic drugs by relieving some of the unpleasant side effects of long-term use or by informing the development of designer medicines. Furthermore, research on younger, high-risk, and prodromal populations may one day make preventive screening and intervention a standard practice in childhood health care (Malhi & Lagopoulos, 2008; Peterson, 2003). All of this is in the future but is founded on the intermediate steps we take now. Through careful consideration of the strengths and weaknesses in neuroimaging, more comprehensive training, and greater intradisciplinary and interdisciplinary communication, we may yet reach the promises of neuroimaging.

Future Directions

There are several challenges whose resolutions will be influential in securing the future of clinical neuroimaging: ensuring proper training, establishing intermediate steps toward incorporating neuroimaging into clinical practice, wrestling with localized and whole-brain activity, and ensuring the proper dissemination of neuroimaging knowledge. The first is important in light of the interdisciplinary nature of the technology and the research in which it is employed. With such a demand to understand physics, biostatistics, clinical neuroscience, and psychopathology among others, how can we establish proper training? The second item is crucial in creating a scientific foundation for any measurements or techniques employed in a clinical setting. The third is a caution: by emphasizing the localization over whole-brain activity, are we blinding ourselves to critical information? Finally, it is evident that neuroimaging is embroiled in "media hype" that can be exploited to make unsound claims and trick patients and their families into unsound practices. What role will clinical scientists play in ensuring that the average person understands the limits to neuroimaging?

References

Arndt, S., Cohen, G., Alliger, R., Swayze, V., & Andreasen, N. (1991). Problems with ratio and proportion measures of imaged cerebral structures. *Psychiatry Research*, *40*, 79–89.

Ashburner, J., & Friston, K. (2000). Voxel-based morphometry—The methods. *Neuroimage*, *11*, 805–821.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: American Psychiatric Publishing Inc.

Bansal, R., Gerber, A. J., & Peterson, B. S. (2008). Brain morphometry using anatomical magnetic resonance imaging. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(6), 619–621.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Beauregard, M. (2009). Effect of mind on brain activity: Evidence from neuroimaging studies of psychotherapy and placebo effect. *Nordic Journal of Psychiatry*, *63*(1), 5–16. doi:10.1080/08039480802421182

Bengtsson, S. L., Nagy, Z., Skare, S., Forsman, L., Forssberg, H., & Ullén, F., et al. (2005). Extensive piano practicing has regionally specific effects on white matter development. *Nature Neuroscience*, *8*(9), 1148–1150. doi:10.1038/nn1516

Bennett, C., Balnd, A., Miller, M., & Wolford, G. (2010). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, *1*(1), 1–5.

Bremner, J., Randall, P., Scott, T., Bronen, R., Seibyl, J., Southwick, S., . . . Innis, R. B. (1995). MRI-based measurement of hippocampal volume in patients with combat-related posttraumatic stress disorder. *American Journal of Psychiatry*, *152*(7), 973.

Brody, A. L., Saxena, S., Schwartz, J. M., Stoessel, P. W., Maidment, K., & Phelps, M. E., et al. (1998). FDG-PET predictors of response to behavioral therapy and pharmacotherapy in obsessive compulsive disorder. *Psychiatry Research*, *84*(1), 1–6.

Brody, A. L., Saxena, S., Mandelkern, M. A., Fairbanks, L. A, Ho, M. L., & Baxter, L. R, Jr. (2001). Brain metabolic changes

associated with symptom factor improvement in major depressive disorder. *Biological Psychiatry, 50*, 171–178.

Bryant, R. A., Felmingham, K., Kemp, A., Das, P., Hughes, G., Peduto, A., & Williams, L. (2008). Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder. *Psychological Medicine*, *38*(4), 555–561. doi:10.1017/S0033291707002231

Burger, C., Goerres, G., Schoenes, S., Buck, A., Lonn, A., & von Schulthess, G. (2002). PET attenuation coefficients from CT images: experimental evaluation of the transformation of CT into PET 511-keV attenuation coefficients. *European Journal of Nuclear Medicine and Molecular Imaging*, *29*(7), 922–927. doi:10.1007/s00259-002-0796-3

Bush, G., Valera, E., & Seidman, L. (2005). Functional neuroimaging of attention-deficit/hyperactivity disorder: A review and suggested future directions. *Biological Psychiatry*, *57*(11), 1273–1284. doi:10.1016/j.biopsych.2005.01.034

Buxton, R., Uludag, K., Dubowitz, D., & Liu, T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage*, *23*(Suppl. 1), S220–S233. doi:10.1016/j.neuroimage.2004.07.013

Buysse, D. J., Hall, M., Begley, A., Cherry, C. R., Houck, P. R., & Land, S., . . . Frank, E. (2001). Sleep and treatment response in depression: new findings using power spectral analysis. *Psychiatry Research*, *103*(1), 51–67.

Chang, L., Cloak, C., & Ernst, T. (2003). Magnetic resonance spectroscopy studies of GABA in neuropsychiatric disorders. *Journal of Clinical Psychiatry*, *64*(Suppl. 3), 7–14.

Courvoisie, H., Hooper, S., Fine, C., Kwock, L., & Castillo, M. (2004). Neurometabolic functioning and neuropsychological correlates in children with ADHD-H: preliminary findings. *Journal of Neuropsychiatry and Clinical Neurosciences*, *16*(1), 63–69.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., . . . Schlaggar, B. L. (2010). Prediction of individual brain maturity using fMRI. *Science*, *329*(5997), 1358–1361. doi:10.1126/science.1194144

Drevets, W. C., Videen, T. O., Price, J. L., Preskorn, S. H., Carmichael, S. T., & Raichle, M. E. (1992). A functional anatomical study of unipolar depression. *Journal of Neuroscience*, *12*, 3628–3641.

Drevets, W. (2001). Neuroimaging and neuropathological studies of depression: Implications for the cognitive-emotional features of mood disorders. *Current Opinion in Neurobiology*, *11*(2), 240–249.

Erwin, R., Gur, R., Gur, R., Skolnick, B., Mawhinney-Hee, M., & Smailis, J., et al. (1992). Facial emotion discrimination: I. task construction and behavioral findings in normal subjects. *Psychiatry Research*, *42*(3), 231–240.

Etkin, A. (2010). Functional neuroanatomy of anxiety: A neural circuit perspective. *Current Topics in Behavioral Neurosciences*, *2*, 251–277.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., . . . Anders, M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Fleck, D. E., Nandagopal, J., Cerullo, M. A., Eliassen, J. C., DelBello, M. P., Adler, C. M., & Strakowski, S. M. (2008). Morphometric magnetic resonance imaging in psychiatry. *Topics in Magnetic Resonance Imaging*, *19*(2), 131–142. doi:10.1097/RMR.0b013e3181808152

Frewen, P. A., & Lanius, R. A. (2006). Neurobiology of dissociation: Unity and disunity in mind-body-brain. *Psychiatric Clinics of North America*, *29*(1), 113–128, ix. doi:10.1016/j.psc.2005.10.016

Fu, C., & McGuire, P. (1999). Functional neuroimaging in psychiatry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *354*(1387), 1359–1370.

Furmark, T., Tillfors, M., Marteinsdottir, I., Fischer, H., Pissiota, A., Langstrom, B., & Fredrikson, M. (2002). Common changes in cerebral blood flow in patients with social phobia treated with citalopram or cognitive-behavioral therapy. *Archives of General Psychiatry*, *59*(5), 425–433.

Genovese, C., Lazar, N., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*(4), 870–870.

Gerber, A. J., & Peterson, B. S. (2006). Measuring transference phenomena with fMRI. *Journal of the American Psychoanalytic Association*, *54*(4), 1319–1325.

Gerber, A. J., & Peterson, B. S. (2008). What is an image? *Journal of the American Academy for Child Adolescent Psychiatry*, *47*(3), 245–248.

Gevins, A., Le, J., Leong, H., McEvoy, L. K., & Smith, M. E. (1999). Deblurring. *Journal of Clinical Neurophysiology*, *16*(3), 204–213.

Goldman, R., Stern, J., Engel Jr, J., & Cohen, M. (2000). Acquiring simultaneous EEG and functional MRI. *Clinical Neurophysiology*, *111*(11), 1974–1980.

Gould, S. (1978). Women's brains. *Natural History*, *87*(8), 44–50.

Gould, S. J. (Ed.). (1981). *The mismeasure of man*. New York: W. W. Norton & Company.

Greenberg, G. (2010, December). Inside the battle to define mental illness. *Wired Magazine*, *19*(2). Retrieved from http://www.wired.com/magazine/2010/ff_dsmv/all/1

Greene, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. doi:10.1126/science.1062872

Gregor, J., & Huff, D. (1997). A focus-of-attention preprocessing scheme for EM-ML PET reconstruction. *IEEE Transaction Medical Imaging*, *16*(2), 218–223.

Gur, R., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., & Macy, L., . . . Gur, R. C. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *Journal of Neuroscience Methods*, *115*(2), 137–143.

Hou, Z. (2006). A review on MR image intensity inhomogeneity correction. *International Journal of Biomedical Imaging*, *2006*, 1–12. doi:10.1155/IJBI/2006/49515

Huettel, S., Song, A., & McCarthy, G. (2004). *Functional magnetic resonance imaging*. Sunderland, CT: Sinauer Associates Inc.

Jackson, G. D., Briellmann, R. S., Waites, A. B., Pell, G., & Abbott, D. (2006). Functional MRI. In G. Webb (Ed.), *Modern magnetic resonance* (pp. 1023–1036). London, England: Springer Academic Press.

Johnston, S. C., Rootenberg, J. D., Katrak, S., Smith, W. S., & Elkins, J. S. (2006). Effect of a US National Institutes of Health programme of clinical trials on public health and costs. *Lancet*, *367*(9519), 1319–1327. doi:10.1016/S0140–6736(06)68578–4

Kaufman, J., & Charney, D. (2000). Comorbidity of mood and anxiety disorders. *Depression and Anxiety*, *12*(Suppl. 1), 69–69.

Kaye, W. (2008). Neurobiology of anorexia and bulimia nervosa. *Physiology & Behavior*, *94*(1), 121–135.

Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., & Thase, M. E., . . . Leon, A. C. (2010). A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry*, *51*(3), 319–324. doi:10.1016/j.comppsych.2009.07.001

Kohler, C. G., Turner, T. H., Bilker, W. B., Brensinger, C. M., Siegel, S. J., Kanes, S. J., . . . Gur, R. C. (2003). Facial emotion recognition in schizophrenia: Intensity effects and error pattern. *American Journal of Psychiatry*, *160* (10), 1768–1774.

Konarski, J., Kennedy, S., Segal, Z., Lau, M., Bieling, P., McIntyre, R., & Mayberg, H. S. (2009). Predictors of nonresponse to cognitive behavioural therapy or venlafaxine using glucose metabolism in major depressive disorder. *Journal of Psychiatry & Neuroscience*, *34*(3), 175.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–883.

Kuo, Y., & Herlihy, A. (2006). Optimization of MRI contrast for pre-clinical studies at high magnetic field. *Modern Magnetic Resonance*, *1*, 759–768.

Le Bihan, D., Breton, E., Lallemand, D., Grenier, P., Cabanis, E., & Laval-Jeantet, M. (1986). MR imaging of intravoxel incoherent motions: Application to diffusion and perfusion in neurologic disorders. *Radiology*, *161*(2), 401–407.

Lehto, S. M., Tolmunen, T., Joensuu, M., Saarinen, P. I., Valkonen-Korhonen, M., Vanninen, R., . . . Lehtonen, J. (2008). Changes in midbrain serotonin transporter availability in atypically depressed subjects after one year of psychotherapy. *Progress in Neuropsychopharmacology and Biological Psychiatry*, *32*(1), 229–237. doi:10.1016/j.pnpbp.2007.08.013

Lindquist, M. A., & Wager, T. D. (2007). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human Brain Mapping*, *28*(8), 764–784. doi:10.1002/hbm.20310

MacCabe, D. P., & Castle, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, *107*(1), 343–352. doi: 10.1016/j.cognition.2007.07.017

Maier, M. (1995). In vivo magnetic resonance spectroscopy: Applications in psychiatry. *British Journal of Psychiatry*, *167*(3), 299–306. doi:10.1192/bjp.167.3.299

Malhi, G., & Lagopoulos, J. (2008). Making sense of neuroimaging in psychiatry. *Acta Psychiatrica Scandinavica*, *117*(2), 100–117.

Marsh, R., Gerber, A. J., & Peterson, B. S. (2008). Neuroimaging studies of normal brain development and their relevance for understanding childhood neuropsychiatric disorders. *Journal of the American Academy of Child Adolescent Psychiatry*, *47*(11), 1233–1251. doi:10.1097/CHI.0b013e318185e703

Martin, S. D., Martin, E., Rai, S. S., Richardson, M. A., & Royall, R. (2001). Brain blood flow changes in depressed patients treated with interpersonal psychotherapy or venlafaxine hydrochloride: Preliminary findings. *Archives of General Psychiatry*, *58*(7), 641–648.

Mathalon, D., Sullivan, E., Rawles, J., & Pfefferbaum, A. (1993). Correction for head size in brain-imaging measurements. *Psychiatry Research: Neuroimaging*, *50*(2), 121–139.

Mayberg, H., Brannan, S., Mahurin, R., Jerabek, P., Brickman, J., & Tekell, J., . . . Martin, C. C. (1997). Cingulate function in depression: A potential predictor of treatment response. *Neuroreport*, *8*(4), 1057–1061.

Mayberg, H. (1994). Frontal lobe dysfunction in secondary depression. *Journal of Neuropsychiatry and Clinical Neuroscience*, *6*(4), 428.

Mayberg, H. (2003). Modulating dysfunctional limbic-cortical circuits in depression: Towards development of brain-based algorithms for diagnosis and optimised treatment. *British Medical Bulletin*, *65*(1), 193.

Mayberg, H. S, Lozano, A. M, Voon, V., McNeely, H. E, Seminowicz, D., Hamani, C., . . . Kennedy, S. H. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45*(5), 651–660.

Mayberg, H. (2007). Defining the neural circuitry of depression: Toward a new nosology with therapeutic implications. *Biological Psychiatry*, *61*(6), 729–730. doi:10.1016/j.biopsych.2007.01.013

McCloskey, M., Phan, K., & Coccaro, E. (2005). Neuroimaging and personality disorders. *Current Psychiatry Report*, *7*(1), 65–72.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–580. doi:10.1016/j.jesp.2009.01.002

Mechelli, A., Price, C., Friston, K., & Ashburner, J. (2005). Voxel-based morphometry of the human brain: Methods and applications. *Current Medical Imaging Reviews*, *1*(2), 105–105.

Moore, C., Biederman, J., Wozniak, J., Mick, E., Aleardi, M., & Wardrop, M., et al. (2006). Differences in brain chemistry in children and adolescents with attention deficit hyperactivity disorder with and without comorbid bipolar disorder: A proton magnetic resonance spectroscopy study. *American Journal of Psychiatry*, *163*(2), 316.

Nemeroff, C., Kilts, C., & Berns, G. (1999). Functional brain imaging: Twenty-first-century phrenology or psychobiological advance for the millennium? *American Journal of Psychiatry*, *156*(5), 671.

Nolan, H., Whelan, R., & Reilly, R. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. doi:10.1016/j.jneumeth.2010.07.015

Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(13), 5951–5955.

Paulus, M. P. (2008). The role of neuroimaging for the diagnosis and treatment of anxiety disorders. *Depression and Anxiety*, *25*(4), 348–356. doi:10.1002/da.20499

Pavuluri, M., Oconnor, M., Harral, E., & Sweeney, J. (2007). Affective neural circuitry during facial emotion processing in pediatric bipolar disorder. *Biological Psychiatry*, *62*(2), 158–167. doi:10.1016/j.biopsych.2006.07.011

Peterson, B. S. (2003). Conceptual, methodological, and statistical challenges in brain imaging studies of developmentally based psychopathologies. *Developmental Psychopathology*, *15*(3), 811–832.

Peterson, B. S., Bansal, R., Chung, Y. A., Dong, Z., Duan, Y., Gerber, A. J., . . . Wang, Z. (2006). Neuroimaging methods

in the study of childhood psychiatric disorders. In A. Martin & F. R. Volkmar (Eds.), *Lewis's child and adolescent psychiatry: A comprehensive textbook* (pp. 214–234). New York: Lippincott.

Peterson, B. S, Skudlarski, P., Anderson, A. W., Zhang, H., Gatenby, J. C, Lacadie, C. M., . . . Gore, J. C. (1998). A functional magnetic resonance imaging study of tic supression in tourette syndrome. *Archives of General Psychiatry*, *55*(4), 326–333.

Peterson, B. S., Staib, L., Scahill, L., Zhang, H., Anderson, C., & Leckman, J. F., et al. (2001). Regional brain and ventricular volumes in Tourette syndrome. *Archives of General Psychiatry*, *58*(5), 427–440.

Pizzagalli, D., Pascual-Marqui, R., Nitschke, J., Oakes, T., Larson, C., & Abercrombie, H., . . . Davidson, R. J. (2001). Anterior cingulate activity as a predictor of degree of treatment response in major depression: evidence from brain electrical tomography analysis. *American Journal of Psychiatry*, *158*(3), 405–415.

Plessen, K., Royal, J., & Peterson, B. (2007). Neuroimaging of tic disorders with co-existing attention-deficit/hyperactivity disorder. *European Child & Adolescent Psychiatry*, *16*, 60–70.

Radua, J., van den Heuvel, O., Surguladze, S., & Mataix-Cols, D. (2010). Meta-analytical comparison of voxel-based morphometry studies in obsessive-compulsive disorder vs. other anxiety disorders. *Archives of General Psychiatry*, *67*(7), 701–711. doi:10.1001/archgenpsychiatry.2010.70

Rapoport, J. L., & Gogtay, N. (2010). Childhood onset schizophrenia: Support for a progressive neurodevelopmental disorder. *International Journal of Developmental Neuroscence*. Advance online publication. doi:10.1016/j.ijdevneu.2010.10.003

Reiman, E. M., Fusselman, M. J., Fox, P. T., & Raichle, M. E. (1989). Neuroanatomical correlates of anticipatory anxiety. *Science*, *243*(4894 Pt 1), 1071–1074.

Ritter, P., & Villringer, A. (2006). Simultaneous EEG–fMRI. *Neuroscience & Biobehavioral Reviews*, *30*(6), 823–838. doi:10.1016/j.neubiorev.2006.06.008

Roffman, J. L., & Gerber, A. J. (2008). Neural models of psychodynamic concepts and treatments: Implication for psychodynamic psychotherapy. In R. Levy & J. S. Ablon (Eds.), *Handbook of evidence-based psychodynamic psychotherapy* (pp. 305–338). Totowa, NJ: Humana Press.

Rykhlevskaia, E., Gratton, G., & Fabiani, M. (2008). Combining structural and functional neuroimaging data for studying brain connectivity: A review. *Psychophysiology*, *45*(2), 173–187. doi:10.1111/j.1469–8986.2007.00621.x

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica: International Journal of Psychonomics*, *30*, 276–315.

Saxena, S., Gorbis, E., O'Neill, J., Baker, S. K., Mandelkern, M. A., & Maidment, K. M., . . . London, E. D. (2009). Rapid effects of brief intensive cognitive-behavioral therapy on brain glucose metabolism in obsessive-compulsive disorder. *Molecular Psychiatry*, *14*(2), 197–205. doi:10.1038/sj.mp.4002134

Sheline, Y. (2003). Neuroimaging studies of mood disorder effects on the brain. *Biological Psychiatry*, *54*(3), 338.

Siegle, G. J., Carter, C. S., & Thase, M. E. (2006). Use of fMRI to predict recovery from unipolar depression with cognitive behavior therapy. *American Journal of Psychiatry*, *163*(4), 735.

Strother, S. (2006). Evaluating fMRI preprocessing pipelines. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 27–41.

Székely, G., Kelemen, A., Brechbühler, C., & Gerig, G. (1996). Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models. *Medical Image Analysis*, *1*(1), 19–34. doi:10.1016/S1361-8415(01)80003-7

Thebérge, J. (2008). Perfusion magnetic resonance imaging in psychiatry. *Topics in Magnetic Resonance Imaging*, *19*(2), 111.

Thomason, M., & Thompson, P. (2011). Diffusion tensor imaging, white matter, and psychopathology. *Annual Review of Clinical Psychology*, *7*(1), 63–85. doi:10.1146/annurev-clinpsy-032210–104507

Tuch, D. S. (2004). Q-ball imaging. *Magnetic Resonance in Medicine*, *52*(6), 1358–1372. doi:10.1002/mrm.20279

Tuch, D. S., Reese, T. G., Wiegell, M. R., Makris, N., Belliveau, J. W., & Wedeen, V. J. (2002). High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine*, *48*(4), 577–582. doi:10.1002/mrm.10268

Westbrook, C., & Kaut, C. (2000). *MRI in practice* (2nd ed.). Oxford, England: Wiley Blackwell.

Wolkin, A., Sanfilipo, M., Wolf, A., Angrist, B., Brodie, J., & Rotrosen, J., et al. (1992). Negative symptoms and hypofrontality in chronic schizophrenia. *Archives of General Psychiatry*, *49*(12), 959–965.

Woods, R., Mazziotta, J., & Cherry, S. (1993). MRI-PET registration with automated algorithm. *Journal of Computer Assisted Tomography*, *17*(4), 536–546.

Wykes, T., Brammer, M., Mellers, J., Bray, P., Reeder, C., & Williams, C., Corner, J. (2002). Effects on the brain of a psychological treatment: Cognitive remediation therapy: Functional magnetic resonance imaging in schizophrenia. *British Journal of Psychiatry*, *181*, 144–152.

Zucker, N. L., Losh, M., Bulik, C. M., LaBar, K. S., Piven, J., & Pelphrey, K. A. (2007). Anorexia nervosa and autism spectrum disorders: Guided investigation of social cognitive endophenotypes. *Psychological Bulletin*, *133*(6), 976–1006. doi:10.1037/0033–2909.133.6.97

CHAPTER

11 Experience Sampling Methods in Clinical Psychology

Philip S. Santangelo, Ulrich W. Ebner-Priemer, *and* Timothy J. Trull

Abstract

The Experience Sampling Method (ESM) can improve our understanding of how psychopathological symptoms unfold over time in everyday life. We discuss major benefits of ESM by presenting selected studies involving (a) real-time assessment (i.e., assessments focusing on individuals' momentary states, experiences, or behaviors); (b) real-world assessments enhancing laboratory-to-life generalizability; (c) multiple assessments over time allowing the study of dynamic processes; (d) multimodal assessment integrating psychological, physiological, and behavioral data; (e) assessment of setting or context specificities allowing for context-sensitive analyses; and (f) the provision of immediate interactive feedback. Furthermore, we offer recommendations concerning design issues for ESM studies, namely with regard to (a) choosing a sampling strategy, (b) participants' burden, compliance, and reactivity, (c) hardware and software solutions, (d) mathematical procedures when analyzing ESM data, and (e) visualization of ESM data. Regardless of remaining challenges, ESM offers great potential in clinical psychology with its possible application as a therapeutic tool and by revealing a comprehensive and generalizable picture of patients' and research participants' symptomatology.

Key Words: Ambulatory assessment, ecological momentary assessment, experience sampling method, e-diary, clinical psychology, psychiatry

Introduction

Contemporary assessment within clinical psychology is dominated by methods that rely on retrospective self-reports of patients. In methods such as unstructured clinical interviews, structured interviews, and self-report questionnaires, patients must recall information about behavioral, emotional, or cognitive symptoms from their memory. However, cognition and memory research has demonstrated that gathering information retrospectively is susceptible to multiple systematic distortions (Ebner-Priemer & Trull, 2009b; Fahrenberg, Myrtek, Pawlik, & Perrez, 2007; Kihlstrom, Eich, Sandbrand, & Tobias, 2000; Stone, Shiffman, Atienza, & Nebeling, 2007) as people rely on so-called memory heuristics. Many memory heuristics have been discussed, including (a) the *mood congruent memory effect,* in which the higher congruency of the current emotional state with the emotional content of information leads to easier retrieval of this information and (b) the *peak-end rule*, which denotes that recall is mainly influenced by peak experience in terms of arousal as well as the most recent experience (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Kihlstrom et al., 2000). The problem with memory heuristics is that they both (a) increase the inaccuracy of retrieved information and therefore increase error variance and (b) constitute systematic errors. Contrary to error variance, systematic errors cannot be countered by simply increasing the number of subjects; thus, the risk of drawing incorrect conclusions and misinterpretations is elevated.

Based on the knowledge gained in the field of memory research, the U.S. Food and Drug

Administration (FDA) released guidelines on the use of patient-reported outcome measures (i.e., patient self-reports of symptomatology) in medical product development, noting that retrospective reports may be invalid due to biases and distortions. The FDA recommends focusing on data assessed in real time (i.e., asking patients to describe their current or very recent state) instead of retrospective assessments (FDA, 2009).

The Experience Sampling Method (ESM; Csikszentmihalyi & Larson, 1987) is a method suited to avoid retrospective biases, as it assesses phenomena of interest in real time or at least as proximal to the actual occurrence as possible. Although different terms have been used for this kind of assessment, namely Ambulatory Assessment (Fahrenberg, Hüttner, & Leonhart, 2001; Fahrenberg & Myrtek, 1996; Fahrenberg et al., 2007), Ecological Momentary Assessment (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002), or Real-Time Data Capture (Stone et al., 2007), all of these methods are characterized by the use of methodology (often computer-based) to repeatedly assess self-reported symptoms, behaviors, or physiological processes while study participants undergo daily life activities. For the sake of simplicity and because we consider these different labels to be interchangeable, we use the label *ESM*.

In addition to overcoming limitations of reliance on retrospection, practitioners and researchers are commonly interested in the symptomatology of a disorder and its associated functional impairments as specifically expressed in real-world environments. Since a variety of studies have shown that laboratory findings do not always translate to real life (Fahrenberg et al., 2007; Horemans, Bussmann, Beelen, Stam, & Nollet, 2005; Wilhelm & Grossman, 2010), examination of experiences, attitudes, or behaviors in everyday life is indispensable. Furthermore, although efficient, the predominant use of global or summary measures hinders the examination of dynamic processes in experiences and behavior over time and across situations.

ESM meets these demands as it is characterized by (a) assessments that focus on individuals' current or very recent states, experiences, or behaviors (i.e. real-time assessment in contrast to retrospective self-assessments demanded by questionnaire and interview methods); (b) collection of data in typical life settings (i.e., real-world environments in contrast to artificial environments like laboratory settings), enhancing real-life generalizability; (c) multiple assessments over time, allowing for investigating time series and studying dynamic processes; (d) the possibility of multimodal assessment integrating psychological, physiological, and behavioral data; (e) assessment of setting or context specificities to reveal associations between symptomatology and context; and (f) possible therapeutic applications by giving interactive feedback in real time and real life. We now consider ESM studies conducted in different fields of clinical psychology, illustrating in turn these six major benefits of the ESM approach to clinical psychology.

Characteristics and Advantages of ESM

Real-Time Assessments: Avoiding Biased Recollection and Reconstruction of the Past

To reduce biases known to distort retrospective reporting and reconstruction of past events and experiences, ESM seeks a more proximal assessment of the experiences, attitudes, or behaviors of interest than is provided by self-report questionnaire or interview measures. Therefore ESM minimizes the risk of and thus the influence of recall biases. Here, we briefly present selected studies that demonstrate the importance and necessity of real-time assessment in investigating symptomatology of panic disorder and borderline personality disorder.

PANIC DISORDER

Several studies have demonstrated retrospective biases in patients' reports of symptomatology when assessed by traditional retrospective assessment tools and compared to ESM data. One of the first studies evaluating retrospective bias in reports of symptomatology is the ESM study by Margraf and colleagues (1987). The authors investigated the phenomenology of panic attacks during daily life in patients with panic attacks (diagnosed according to DSM-III criteria; American Psychiatric Association, 1980) relative to healthy controls (HCs). Participants completed a panic diary (i.e., the Stanford Panic Attack Diary, a modified version of the diary used by Taylor et al., 1986) for 6 consecutive days in which the occurrence of a panic attack and accompanying symptoms were reported. During 6 days, patients reported a total of 175 panic attacks. In addition to the ESM assessment, the authors retrospectively assessed the phenomenology of patients' prototypical panic attacks using a disorder-specific questionnaire and a structured diagnostic interview (i.e., Structured Clinical Interview for DSM-III—Upjohn Version; Spitzer & William, 1983). Real-time and retrospective reports of symptom patterns offered very different clinical portrayals. Specifically, there

was a remarkable discrepancy between the number of symptoms reported by the two retrospective assessment methods and by the diary measure. The most striking difference concerned the symptom "fear of dying." This was reported in 70 percent of panic attacks in the questionnaire but in only 3 percent of the panic attacks in the diary. The same applied to the symptom "faintness" as it occurred in 89 percent of panic attacks according to the questionnaire assessment but in only 10 percent of the panic attacks according to the diary. Such retrospective exaggeration of symptoms was apparent for all panic symptoms. Furthermore, patients reported in the ESM diary that 3 panic symptoms occurred in an average panic attack; in the retrospective questionnaire they reported 11 symptoms. Although the retrospective methods did not refer to the same specific panic attacks experienced during the diary assessment period, the differences between the cognitive representation and the actual characteristics of panic attacks seem to be considerable.

De Beurs and colleagues (1992) replicated these findings. They compared reported frequencies of panic attacks in patients with a diagnosis of panic disorder with agoraphobia according to (a) ESM (using a daily monitoring approach; i.e. patients filled out paper-and-pencil forms describing DSM-III-R panic attack symptoms after experiencing a panic attack) and (b) retrospective estimations using the Mobility Inventory (Chambless, Caputo, Jasin, Gracely, & Williams, 1985), which requires patients to specify the occurrence of panic attacks during the past week. Comparisons of reports over 12 weeks revealed a retrospective overrating of panic attack frequency, especially at the beginning of the assessment period. There were tremendous discrepancies between the two methods in the reported frequencies of panic attacks: the mean panic frequency according to ESM was 0.97 panic attacks per week compared to 1.93 using the retrospective questionnaire measure.

BORDERLINE PERSONALITY DISORDER

To examine retrospective bias in patients with borderline personality disorder (BPD), Ebner-Priemer and colleagues (2006) compared momentary mood ratings collected every 10 to 20 minutes over 24 hours by ESM with single retrospective mood ratings over the same period in patients with BPD and in HCs. Both groups showed a valence-dependent recall bias, but in opposite directions. Reports of HCs revealed a retrospective overestimation of the intensity of positive moods and retrospective underestimation of the intensity of negative moods. In contrast, BPD patients were more likely to retrospectively overestimate the intensity of moods with a negative valence and to underestimate the intensity of moods with a positive valence.

In another example, Solhan, Trull, Wood, and Jahng (2009) examined the concordance of retrospective self-reports with momentary assessments of affective instability in outpatients with BPD or depressive disorder. For momentary assessments, the authors used repeated assessments of mood collected via e-diaries at six random times per day over a 28-day period. For retrospective trait measurement of affective instability, the Affective Instability subscale of the Personality Assessment Inventory—Borderline Features Scale (Morey, 1991), the Affect Lability Scales (Harvey, Greenberg, & Serper, 1989), and the Affect Intensity Measure (Larsen, Diener, & Emmons, 1986) were administered immediately following the 28-day ESM assessment period. Results revealed that there was almost no relationship between retrospective questionnaire trait measures and ESM indices of affective instability; the relationships were modest at best and not at all satisfactory. In addition, Solhan and colleagues (2009) examined whether patients with BPD could at least remember their most pronounced mood changes (defined as the highest 10 percent of change scores across all participants on a particular scale). From a memory heuristic perspective, single important events (i.e., peaks) should be remembered more easily (a phenomenon referred to as the peak-end rule; Kahneman et al., 1993). However, retrospective reports of extreme mood changes were largely unrelated to ESM indices of acute affect changes, regardless of whether the previous month or the immediately preceding 7 days were addressed.

These findings of Solhan and colleagues (2009) highlight the special problem inherent in the retrospective estimation of "unstable" symptoms. Ebner-Priemer and colleagues (2007a) examined concordance between ESM data and expert interview ratings for two BPD criteria: (a) the stable averaged experience criterion "inappropriate, intense anger" and (b) the, by definition, unstable criterion "affective instability." Results showed at least some concordance between ESM measures and expert ratings when investigating the stable criterion "inappropriate anger," but no concordance between the two methods was found in the assessment of affective instability.

Taken together, these studies underscore problems in the recall of symptoms as required in

questionnaires and interviews. Retrospective exaggeration and overestimation of disorder-specific symptoms may be the rule rather than an exception, as it has also been shown regarding both negative and positive affect in depression (Ben Zeev & Young, 2010; Ben Zeev, Young, & Madsen, 2009), bingeing and excessive exercise in eating disorders (Stein & Corte, 2003), obsessions and compulsions in obsessive-compulsive disorder (Herman & Koran, 1998), catastrophic cognitions in agoraphobics (Marks & Hemsley, 1999), and pain intensity ratings in pain patients (Stone & Broderick, 2007).

ESM may also provide more accurate assessments of the economic effects of mental health problems in comparison to traditional calculation methods (i.e., estimates usually based solely on the amount of days absent from work). Wang and colleagues (2004) investigated the effects of major depression on work performance in 105 airline reservation agents and 181 telephone customer service representatives using ESM methodology. The authors used paper-and-pencil diaries and a pager to assess the momentary work performance (by asking questions regarding task focus [i.e., questions about the ability to concentrate and focus on work] and productivity [i.e., questions about quality, speed, and efficiency]) at five random times each day over 7 days. Depressed workers, although attending work, reported impaired performance and reduced work productivity in their diaries, with effect sizes corresponding to a 0.4 standard deviation decrease in task focus and a 0.3 standard deviation decrease in productivity. This finding suggests that studies focusing only on the amount of days absent from work may significantly underestimate the economic effects of depression, since—although attending work—depressed workers were far less productive.

Even though the examples above indicate a potentially higher accuracy of ESM compared to traditional assessment methods, it is not evident which method should be used to assess symptoms. From a clinical standpoint it might be more informative to know a patient's retrospective evaluation of his or her symptoms. However, the problem with retrospective self-report measures is that they may not only be exaggerated, but can also be greatly influenced by a subject's current context and momentary mental state (Fredrickson, 2000; Kahneman et al., 1993; Kihlstrom et al., 2000). Consider the mood-congruent memory effect. While in a good mood, a patient with panic disorder might have more difficulties remembering negative symptoms of a panic attack, whereas in a bad or anxious mood the retrieval may be enhanced. Thus, a person's current state and situation at the moment of reporting will determine or at least influence what is reported. Therefore, the recall of memories is not stable—a tendency that can cause grave methodological problems (Stone & Broderick, 2007). However, very recent considerations go beyond the assumption that ESM data are generally more accurate and thus superior to retrospective self-report measures assessed by questionnaire or interview procedures (Conner & Feldman Barrett, 2012). The authors offer recommendations for choosing the most appropriate self-report procedure for a particular research question, based on findings that, depending on the measure used to obtain self-reports, different types of "self" (i.e. the *experiencing*—momentary, *remembering*—retrospective, and *believing*—trait self) are assessed (which in turn are functionally and neuroanatomically different). Depending on the research hypotheses, one has to select the self-report procedure best suited for one's own particular research question.

Enhancing Generalizability: Assessment in Real-Life Situations

Real-life assessment is one of the most distinct advantages of ESM, as symptoms are investigated where they actually occur and where patients suffer from them: in patients' everyday lives. Whereas laboratory studies offer the possibility of testing hypotheses under the most rigorous control, they nonetheless do so under artificial, laboratory conditions. This may adversely affect construct, ecological, and external validity, accounting for differences between the laboratory and real life (Fahrenberg et al., 2007; Horemans et al., 2005; Wilhelm & Grossman, 2010).

The phenomenon of office hypertension, also called the "white-coat effect," is the most impressive example showing that phenomenology inside and outside the laboratory may differ. This effect is defined as the occurrence of heightened blood pressure when measured in the medical environment, whereas a subject's blood pressure in everyday life is within the normal range. This phenomenon has been supported in a multitude of studies. Even though patients with "office" hypertension are at relatively low risk of cardiovascular morbidity (Verdecchia et al., 1998), this is of significant practical importance since it has direct implications for diagnosis and treatment. Because of this effect, thousands of people are misdiagnosed every year and unnecessarily medicated (see Hansen, Jeppesen, Rasmussen, Ibsen, & Torp-Pedersen, 2006). Moreover, the

importance of this effect becomes obvious when looking at the prognostic value of blood pressure readings. For example, Salles and colleagues (2008) supported the supremacy of ambulatory blood pressure in predicting cardiovascular morbidity and mortality; office blood pressure did not show any prognostic value.

This serves as an instructive reminder of how potentially fallible it may be to generalize solely on the basis of laboratory experiments and "clinic" assessments, which may be biased for a number of reasons. This example from another discipline—internal medicine—raises the pertinent question of whether missing empirical data of daily life may have led to systematic misinterpretations in clinical psychology as well.

As ESM assesses experiences, attitudes, or behaviors in the contexts in which they naturally occur, such data are assumed to be construct, ecologically, and externally valid. Nonetheless, we view ESM and laboratory methods not as fundamentally opposed alternatives, but instead, by enhancing the laboratory-to-life generalizability of findings to real-world, real-life experience, we think that ESM provides a valuable additional approach to laboratory studies and in-office assessments. Moreover, ESM must empirically evidence its superior validity over laboratory studies, and this is still lacking right now.

Repeated Assessment: Investigating the Variability of Experience and Within-Person Processes

Frequent, repeated assessments are a defining characteristic of ESM. The time series and the timely resolution afforded by multiple measures offer a detailed picture of constructs with dynamic nature. This enables researchers to investigate dynamics of experience and how symptoms vary over time and across contexts by identifying within-subject processes and the dynamic interplay among environmental factors, personal experiences, and psychopathological symptoms. Therefore, ESM is an especially pertinent investigation tool for clinical disorders defined by unstable or cyclic patterns of mood, such as bipolar disorder or BPD (Ebner-Priemer, Eid, Kleindienst, Stabenow, & Trull, 2009).

Because affective instability is a defining characteristic of BPD (American Psychiatric Association, 2000; World Health Organisation, 1992), portraying the dynamic process of transient, fluctuating affective states, the use of ESM is especially applicative. Ebner-Priemer and colleagues (2007b) used electronic dairies and a high sampling frequency (every 10 to 20 minutes during a 24-hour period of everyday life) to assess current affective states of patients with BPD and HCs. Participants were asked to report the occurrence and intensity of current affective states and current intensity of distress. Results show heightened affective instability for both emotional valence and distress in the BPD group in contrast to the HC group. Additionally, the BPD group showed a group-specific pattern of instability characterized by rapid fluctuations from a mood state with positive valence to a mood state with negative valence. Such pronounced and abrupt mood changes correspond with the clinical portrait of BPD patients.

Trull and colleagues (2008) used electronic diaries to record current affective states (using items from the Positive and Negative Affect Schedule [PANAS], as well as its expanded form, the PANAS-X; Watson & Clark, 1999) six times a day over a 28-day period in patients with BPD and patients with a depressive disorder. Results revealed no differences between the BPD group and the depressive disorder group regarding mean levels of positive and negative affect. However, patients with BPD experienced more instability in hostility, fear, and sadness compared to the control group. Extreme changes across successive occasions (defined as change scores greater than or equal to the 90th percentile of the total successive difference across all participants on a particular scale) were also more frequent for hostility in BPD patients. Also noteworthy is the unexpectedly high instability of mood states in the depressed patients: When graphically visualizing single time courses of raw negative-affect scores for a study participant of the BPD group and of the depressive group, it is difficult to distinguish which course of negative affect belongs to which patient group (Fig. 11.1).

A recent study also revealed heightened affective variability in depression (Peeters, Berkhof, Delespaul, Rottenberg, & Nicolson, 2006), leaving justifiable doubt about the "stability" of symptoms thought to be (relatively) stable, like depressive affect. Such symptoms may actually show a significant amount of variability over time when assessed via time-sensitive methods like ESM. In addition, repeated assessments to track affective symptoms over time seem necessary if not obligatory, as the congruence between retrospective assessments of instability and the actual ebb and flow of symptoms is moderate at best (Solhan et al., 2009).

ESM offers another valuable area of application; it allows the investigation of the antecedents and consequences of experiences, attitudes, or behaviors

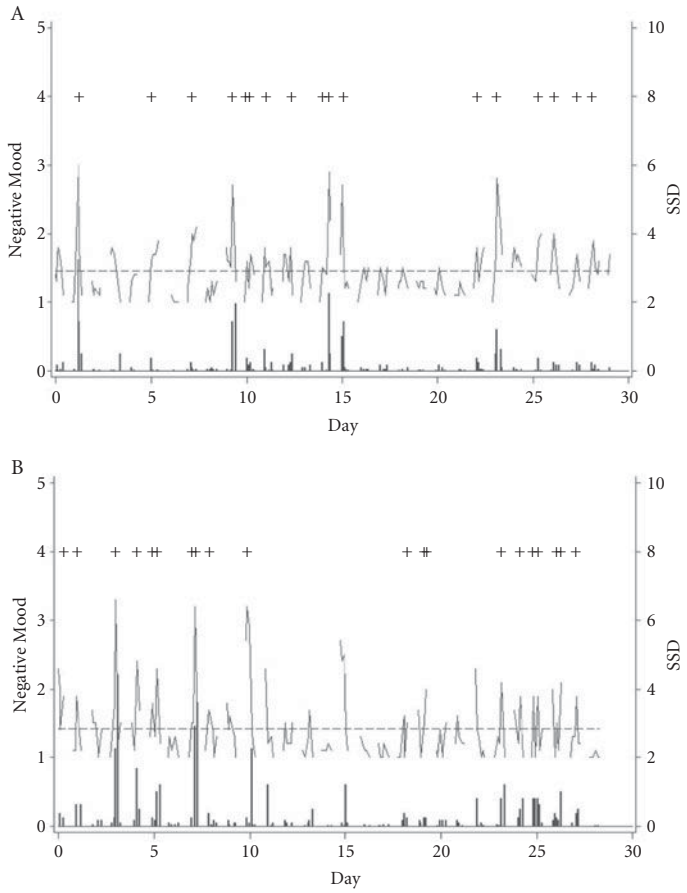


Figure 11.1 Raw scores of negative affect over the course of the 28-day assessment period for (**A**) a patient in the major depressive disorder group (157 assessment points) and (**B**) a patient in the BPD group (156 measurement occasions). Solid lines represent raw scores of negative affect (gaps between solid lines indicate between-day assessments); the dashed line represents the mean score of negative affect over all assessment points; plus symbols represent extreme changes across successive occasions (defined as change scores greater than or equal to the 90th percentile of the total successive difference across all participants on a particular scale); and bars represent instability of negative affect as they indicate the squared successive differences (SSD) between the current and the previous occasions. Differentiation of the two groups (BPD vs. major depressive disorder) is not clearly visible to the naked eye.

(e.g., the antecedents of dysfunctional behavior). Because the antecedents are assessed before the dysfunctional behavior occurs, the assessment of the state before the dysfunctional behavior is not biased by the dysfunctional behavior itself. This distinguishes ESM from traditional questionnaire or interview techniques where assessments of the "typical" phenomenology necessitate simultaneous reports of the state before and after dysfunctional behavior.

For example, Smyth and colleagues (2007) investigated the emotional antecedents and consequences of binge/purge behavior in female patients with bulimia nervosa. E-diaries were used to prompt patients six times a day for 2 weeks to answer questions regarding affective state and binge/purge behavior. Multilevel analyses revealed that a decrease in positive affect and an increase in negative affect and anger/hostility preceded binge/purge episodes in the flow of daily life. Conversely, patients reported an increment of positive and a decrement of negative affect in the aftermath of binging/purging behavior. These findings support an emotion-regulatory function of dysregulated eating behaviors as a maladaptive attempt to alleviate negative affect. Similar findings have been obtained regarding the effects of nonsuicidal self-injury acts in bulimia nervosa. Using the same dataset as Smyth and colleagues (2007), Muehlenkamp and colleagues (2009) investigated the affective antecedents and consequences of nonsuicidal self-injuries. Findings partially support an emotion-regulation

model of self-injurious behavior since decreasing positive affect and increasing negative affect were reported before acts of self-injuries, whereas positive affect increased after acts of self-injuries (while negative affect remained unchanged).

However, the preliminary results of our ongoing study on the effect of dysfunctional behavior on affect and tension in patients with BPD reveals a less consistent picture (Santangelo, Ebner-Priemer, Koudela, & Bohus , 2010). In BPD dysfunctional behaviors in general (like high-risk behavior, binging/purging, substance use, sexual impulsivity, or—the most prominent dysfunctional behavior in BPD—self-injuries) are commonly seen as maladaptive coping strategies used to gain relief from painful negative affect and aversive states of high tension. However, the data so far present a different picture. For a single BPD patient, Figure 11.2 depicts the course of tension (dark-gray dotted line) and valence (light-gray dotted line) over one day. Furthermore, bars represent reports of dysfunctional behaviors. Each bar marks a time frame of approximately 1 hour in which the dysfunctional behavior occurred. Due to the coding of tension and valence, an increase equates to an improvement of the momentary affective state (i.e., a decline in tension and an amelioration in valence); decreases in tension and valence equate to a worsening of the affective state (i.e., an increase in tension and a worsening in valence).

Contrary to the expectations, dysfunctional behaviors were not always followed by a decrease in tension and an improvement of affective state. Instead, there was no general pattern, as sometimes dysfunctional behavior was followed by a worsening of the affective state. Overall, the data reveal that patients show "cascades" of different dysfunctional behaviors. Based on the dataset consisting of 22 patients with BPD, assessed in hourly intervals for 4 consecutive days, we speculate that intermittent reinforcement might be a possible explanation for the maintenance of dysfunctional behaviors even if they do not always help to regulate negative emotional states.

In sum, ESM aims to assess the dynamics of experiences, attitudes, or behaviors over time in everyday life, capturing life as it is lived. Furthermore, it seeks to clarify the affective, cognitive, or behavioral circumstances (i.e., the antecedents and consequences) under which a certain behavior of interest occurs. Because ESM provides time series data, it is well suited for these kinds of research questions and has offered a wide range of insights in various fields of clinical psychology (see also Ebner-Priemer & Trull, 2009b). Several recent reviews illustrate the successful use of ESM and its timely resolution to address various research objects—such as in the field of substance use (Shiffman, 2009), mood disorders (Ebner-Priemer & Trull, 2009a), psychosis (Oorschot, Kwapil, Delespaul, & Myin-Germeys, 2009), and health psychology in general (Smyth & Heron, 2011).

Multimodal Assessment: Integrating Psychological, Physiological, and/or Behavioral Data

Although self-reports assessed by ESM may be more reliable and valid than reports relying on

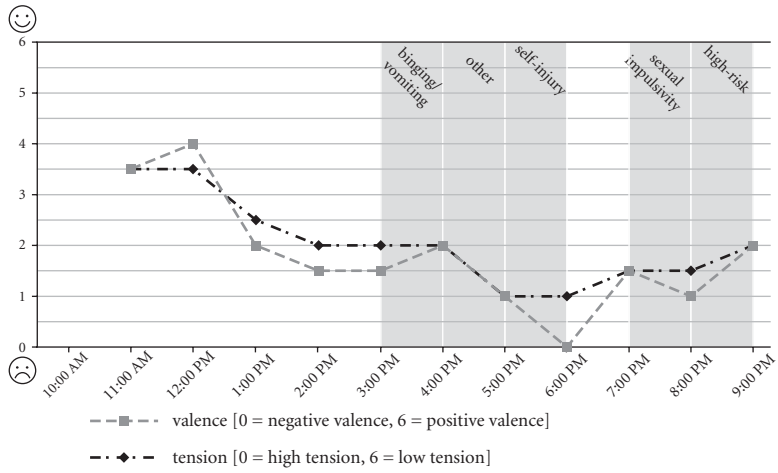


Figure 11.2 The course of tension and valence over one assessment day as well as reports of dysfunctional behaviors. Coding: A high score is associated with a positive affective state on both scales (i.e., a high score on tension is equal to a state of low tension, and a high score on valence is equal to a positive valence). Bars mark reports of dysfunctional behaviors. Contrary to expectations, dysfunctional behaviors were not always followed by a decrease in tension and an increase in positive affect.

retrospection, ESM nonetheless relies on a subjective evaluation by the patient. Therefore, it may be desirable to supplement these psychological measures with objective measurements of physiological and/or behavioral data. Since the advent of mobile high-capacity microsensors, ESM comprises the assessment of not only psychological but also physiological and behavioral data. This is very attractive because multiple studies have shown discrepancies between recalled self-reports of symptoms and objective measures of symptoms. For example, reviews focusing on the congruency between subjective (i.e., self-reports, such as questionnaires) and objective (i.e., directly measured, such as via accelerometry) measures of physical activity indicate a low to moderate agreement both in adult populations (Prince et al., 2008) and pediatric populations (Adamo, Prince, Tricco, Connor-Gorber, & Tremblay, 2009). Although no clear pattern for the differences between self-report and objective measures has been established in adult populations, a method-dependent overestimation of physical activity measures appears evident in pediatric populations (with subjective self-report measures overestimating the objectively measured values). Therefore, adding objective assessments of the symptoms of interest provides a more complete picture of the symptomatology as it unfolds in patients' everyday habitats (Bussmann, Ebner-Priemer, & Fahrenberg, 2009).

An instructive example for the use of sophisticated biosensor technology is the study by Wilhelm and Roth (1998) with the apt title "Taking the laboratory to the skies." The authors used an ambulatory recorder system to examine individuals with flight phobia and sex- and age-matched HCs during a 12-minute flight in a small turboprop airplane. The assessment included several physiological parameters (obtained from cardiovascular, electrodermal, and respiratory activity measures) as well as self-reports of anxiety, tension, excitement, and a short questionnaire comprising the DSM-III panic attack symptoms. Self-reports were assessed three times: at the preflight baseline, shortly after takeoff, and at the postflight baseline. For the statistical analysis, 120-second periods of the physiological measures at the corresponding self-report measure times were used. As expected, all self-report measures of anxiety and several physiological measures (including heart rate, additional heart rate, respiratory sinus arrhythmia, skin conductance fluctuations, and inspiratory pause) changed more during the flight in the flight-phobics than in HCs. In particular, sympathetic activation in flight-phobics with the *in vivo* introduction of the anxiety-inducing stimuli was enhanced, whereas cardiac parasympathetic activation was reduced. Discriminant analyses showed clear group classifications using either self-report or physiological measures. However, a direct statistical comparison of the effect sizes for heart rate and self-rated anxiety change scores showed that heart rate was significantly poorer in distinguishing groups compared to self-ratings.

Another example for the application of a psychophysiological assessment in patients' natural environment is a study conducted by Lemke and colleagues (1997) that investigated patients with a diagnosis of major depression with melancholic features over 3 days. Actimeters were used to continuously assess physical activity; in addition, patients judged their subjectively experienced intensity of symptoms twice a day, in the morning and in the evening. Patients reported feeling significantly less active and awake and more depressed in the morning compared to the evening. Nevertheless, actigraphically measured motor activity indicated the opposite: there was significantly greater motor activity in the morning compared to evening hours. Thus, motor activity was negatively correlated with subjectively experienced symptom intensity. Even though this result is to be regarded as preliminary since the sample in this study was quite small, it shows that prevailing clinical assumptions should be verified with objective data assessed in everyday life.

Even though changes in physical activity are ubiquitous across psychiatric disorders, studies assessing behavioral activity are relatively uncommon (Tryon, 2006). However, adding objective physiological assessments can reveal new insights into both physiological and behavioral components of psychological disorders in everyday life. Today's biosensor technology offers compact, portable, and unobtrusive recording systems that allow assessment in the field (see Ebner-Priemer & Kubiak, 2007; Kubiak & Krog, 2011; see also "Hardware and Software Solutions," below, for discussion of physiological ambulatory monitoring solutions). Furthermore, sophisticated computer processing enables the control of confounding variables outside the laboratory, such as the disentangling of emotional activation from the activation of physical effort (Houtveen & de Geus, 2009; Intille, 2007). Furthermore, the seminal use of physiology-triggered sampling protocols offers the possibility of examining various new research questions.

Assessing the Context of the Report: Allowing for Investigations of Setting- or Context-Specific Relationships

Traditional assessment approaches, such as symptom questionnaires and interviews, are limited in revealing contextual information since the context itself is not assessed. In contrast, the repeated assessments in ESM offer the possibility of assessing varying contexts and situations. This allows researchers to analyze situational influences on symptomatology (i.e., context-sensitive analyses). Using traditional assessment tools, relevant symptomatology is usually assessed for a certain period of time (e.g., the past week, the past month) but not for specific situations (e.g., while alone, while with others). In ESM, instead, both symptoms and context information can be assessed repeatedly and simultaneously over time. An early example of an ESM study addressing setting-specific symptomatology in patients with mental illness and in nonpsychiatric control subjects was conducted by Delespaul and deVries (1987). The authors used digital wristwatches to randomly signal participants ten times a day for 6 consecutive days to fill out paper-and-pencil booklets assessing, among other things, contextual information (i.e., questions regarding where and with whom they were and what they were doing). Results showed that psychopathological symptoms differed as a function of context, such as being alone, being with others, at home, or out of the house. However—and contrary to expectations—reported day-to-day experience indicated that patients felt as well or better when away from home and with other people than did HCs.

Stiglmayr and colleagues (2008) provided an illustrative example for a context-sensitive examination in patients with BPD. Dissociative symptoms in DSM-IV BPD co-occur with states of intense distress. Therefore, it is expected that dissociative symptoms in BPD would be present during states of high distress but not during states of low to medium distress. Stiglmayr and colleagues (2008) used e-diaries to assess psychological and somatic dissociative symptoms as well as subjective ratings of distress every waking hour for 2 days in BPD patients, clinical control subjects (i.e., those with major depression or panic disorder), and HCs. Distress was associated with dissociation in all groups, but this association was, as hypothesized, most pronounced in BPD patients. Consistent with DSM-IV criteria, BPD patients were more prone to dissociation when experiencing distress. However, the dissociation–stress association was linear (i.e., increased distress was accompanied by increased dissociative symptoms) and thus not only related to periods of extreme distress as stated in DSM-IV.

A final example of a context-sensitive investigation is the assessments of diurnal variation of symptoms, referring to the examination of symptom intensity in relation to the time of day. ESM has been used to investigate diurnal symptom patterns in patients with major depressive disorder (e.g., Peeters et al., 2006; Volkers et al., 2003), in nonclinical individuals with varying levels of depressed mood (Murray, 2007), and in patients with bulimia nervosa (Smyth et al., 2009), among other conditions.

Taken together, ESM appears particularly well suited for context-sensitive assessment and analysis. It can help to clarify whether certain symptoms are elicited by, are maintained by, and/or are the result of specific events or contexts and offers the opportunity of a much more representative and nuanced contextual investigation of psychopathological symptoms.

Interactive Assessment: Giving Real-Time Feedback in Real-World Environments and Real-Life Situations

ESM has great potential as a therapeutic application tool because it offers the possibility of triggering electronically mediated interventions *in situ*. This has been termed "interactive assessment," denoting that the answer given to a current question affects future questions, beeps/prompts, or text statements (Fahrenberg, 1996; Shiffman, 2007). In general, two forms of interactivity in ESM can be distinguished: interactive ESM assessment and interactive ESM assessment with individually tailored moment-specific feedback. In addition, ESM can be extended with treatment components.

INTERACTIVE ESM ASSESSMENT

Perhaps the simplest form of interactive ESM assessment is *branching*. In this case, specific questions are administered only if a predefined response occurs. For example, questions about intoxication might be administered only if a patient endorsed the consumption of a certain number of alcoholic drinks. Branching is often used with the intent to reduce patients' assessment burden by administering only the most relevant items and leaving out unnecessary ones. An example for a simple branching is the study by Stiglmayr, Gratwohl, Linehan, Fahrenberg, and Bohus (2005). The authors used e-diaries to assess aversive tension in hourly intervals over a 2-day period. The e-diaries were programmed

to elicit a conditional question regarding predefined changes in tension, asking patients to report events associated with these changes in tension.

Similar to branching is physiology- or context-triggered sampling (Intille, 2007). This can be seen as a kind of "intelligent" sampling because a specific item or sampling protocol (e.g., psychological assessments such as mood state, cognition, attitude, etc.) is prompted in response to a predefined physiological event (e.g., increase in heart rate) or situational context (e.g., voice of a partner). Myrtek (2004) developed a sophisticated algorithm that signals the participant to make a self-report depending on his or her physiological arousal. Since heart rate and physical activity are measured and compared online, events with high emotionally induced physiological arousal (heart rate increase without increase in physical activity) and events with less physiological arousal (no heart rate increases) can be identified during the ongoing assessment. These "detected" events trigger an e-diary device to signal a self-report request. Myrtek and colleagues have used this interactive physiology-triggered ESM approach to investigate a large number of participants in about 20 studies. For a detailed description of the method and of completed studies, see Myrtek (2004).

INTERACTIVE ESM ASSESSMENT WITH INDIVIDUALLY TAILORED MOMENT-SPECIFIC FEEDBACK

When interactivity of ESM is used not only for branching but also for giving individually tailored moment-specific feedback, the distinction between assessment and treatment becomes almost completely blurred. Immediate feedback can advise patients about how to cope with symptoms while undergoing daily life activities. Thus, the treatment is provided in real time in the real world.

One example of the therapeutic application of ESM is evident in a study conducted by Solzbacher and colleagues (2007) in which they used ESM with individually tailored moment-specific feedback to reduce states of affective dysregulation in patients diagnosed with BPD, chronic posttraumatic stress disorder, and bulimia nervosa. A cell phone tracked patients' symptoms over time by assessing current affective states and states of distress four times a day. If distress exceeded a critical intensity (i.e., a predefined cutoff value), patients automatically received a reminder on how to regulate their distress. These reminders mostly suggested the use of skills from the Dialectical Behavior Therapy skills training (Linehan, 1993), in particular emotion regulation and distress tolerance skills. After 30 minutes, an additional prompt assessed the momentary affective state to examine the usefulness of the advice and skills use. Although the reported findings of this ongoing study are preliminary, they are encouraging, showing the feasibility of ESM to provide automated individually tailored moment-specific feedback.

Another example of ESM with individually tailored moment-specific feedback is the work of Tryon and colleagues (2006), who used actigraphy devices to continuously monitor activity level and motor excess in 8- to 9-year-old boys diagnosed with attention-deficit/hyperactivity disorder (ADHD). The authors used both vibratory feedback and visual feedback regarding current and cumulative activity to reduce activity levels during school periods. Most of the participants reduced their activity level from 20 to 47 percent of baseline levels, while only two participants slightly increased their activity level (one by 2 percent and the other by 7 percent of baseline levels). The difference between laboratory-based biofeedback approaches and the study by Tryon and colleagues (2006) is that the problematic behavior in the ESM approach was directly addressed and modified in daily life, therefore bypassing the potential generalization problem of in-office treatment.

ESM WITH TREATMENT COMPONENTS

Kenardy and colleagues (2003) were interested in the treatment effect of ESM when used to prompt patients to practice therapy components in their natural environment. Therefore, they investigated the cost-effectiveness of a brief (6-session) individual cognitive-behavioral therapy (CBT) treatment supplemented with ESM compared to a brief (6-session) CBT treatment without ESM augmentation and a standard (12-session) CBT treatment in 163 patients with panic disorder. In the group with the ESM supplement, the participants received an e-diary following six CBT sessions. The e-diary automatically signaled participants at five fixed times daily to remind them to practice therapy components. Results show that the symptomatology of all three treatment groups improved compared to wait-list patients. Specifically, treatment outcomes were best for patients in the 12-session CBT group, followed by the 6-session computer-augmented treatment group, and finally by the group receiving 6 sessions without ESM augmentation. Even though 6 sessions of CBT were inferior to 12 sessions of CBT, the use of computer augmentation resulted in

a better outcome compared to the brief treatment without computer augmentation.

Another example is an intervention designed for panic patients that targets respiratory functions (Meuret, Wilhelm, Ritz, & Roth, 2008). Patients with a diagnosis of panic disorder were provided with a portable capnometer device, which analyzes exhaled breath. The intervention consists of educational components as well as capnometry-assisted breathing training exercises to be performed twice daily in the natural environment (i.e., at home or elsewhere outside a clinical context). Thus, patients monitored and addressed respiratory dysregulation as it occurred in their natural environment. Results from 20 patients showed significant improvements with respect to disorder severity, agoraphobic avoidance, anxiety sensitivity, disability, and respiratory measures. This improvement was maintained at a 2-month as well as a 12-month follow-up.

In sum, a key feature of studies using ESM for interventions is that the treatment is provided immediately while participants undergo normal daily life activities (i.e., in real time and in daily life). The most salient advantage of this kind of intervention feedback compared to feedback in a standard treatment session (typically once a week in a therapeutic setting) is that patients directly use therapeutic advice in their natural environment. Therefore, the (serious) problem of generalizing behavior learned in a treatment setting to situations in everyday life is overcome. To date, the superiority of ESM interventions over treatment as usual has not been definitively shown, and the current state of knowledge does not allow for drawing strong conclusions at this time. Nonetheless, mobile technology-based ESM feedback (i.e., via palmtop computers and mobile phones or smartphones) seems to offer a promising adjunct to and augmentation of more traditional interventions. Results so far indicate high feasibility and wide acceptance among study participants.

Method—Issues in Planning and Designing ESM Studies

The major considerations of design and implementation issues for ESM studies are choosing an appropriate sampling strategy, which can influence participants' burden, compliance, and reactivity. We will briefly review existing hardware and software solutions, discuss recommendations on the mathematical procedures when analyzing ESM data, and provide examples of how to visually present ESM data.

First, it is most important to recognize that the research question is the center of every decision when planning an ESM study. The research question clearly determines the adequacy of the sampling strategy, the choice of hardware and software, as well as the data analytic strategy and the graphical data description.

Choosing an Adequate Sampling Design

Traditional approaches such as questionnaires or interviews typically entail single-occasion assessments. In contrast, ESM is characterized by multiple repeated assessments. Therefore, a sampling protocol, the scheme defining the scheduling and temporal coverage of the assessment period, is necessary. Because ESM aims to gain a representative sample of a subject's experience or behavior, the proper design of the ESM protocol is essential and represents the first step. Generally, there are seven sampling strategies for assessing psychophysiological data: continuous monitoring, time- or event-dependent monitoring, ambulatory psychometric testing, field experiment, interactive monitoring, symptom monitoring, as well as combinations of these sampling protocols (Fahrenberg et al., 2007). Here, we focus on three sampling strategies: (a) time-contingent sampling, (b) event-contingent sampling, and (c) combined sampling. In addition, we will discuss interactive sampling—in particular physiology-triggered assessment—as it represents a very sophisticated application of ESM. Comprehensive overviews and descriptions of the different sampling strategies, which are beyond the scope of this chapter, can be found elsewhere (Fahrenberg & Myrtek, 2001; Fahrenberg et al., 2007; Piasecki, Hufford, Solhan, & Trull, 2007; Shiffman, 2007).

CHOOSING A SAMPLING PROTOCOL

The defining characteristic of *time-contingent sampling* protocols is multiple repeated assessments over time (e.g., every hour, randomized assessments within a certain time period). Time-contingent recordings are particularly well suited for examining the dynamics of continuous symptomatology, such as the changes and patterns in manic and depressive symptoms in bipolar disorders over time (Bauer et al., 2006) or investigating affective instability in BPD (Ebner-Priemer et al., 2007b). In contrast, *event-contingent sampling* protocols gather data only when a specific event occurs or under certain context conditions (i.e., data assessment is organized around predefined discrete events). Event-contingent recordings are particularly useful when investigating

dynamic influences, such as the smoking relapse process (Shiffman, 2005) or interpersonal problems in BPD patients (Russell, Moskowitz, Zuroff, Sookman, & Paris, 2007). So-called *combined sampling* protocols integrate the two aforementioned time- and event-contingent approaches. Combined recordings are best used in the examination of the interplay between events and dynamic phenomena, such as in the investigation of antecedents and consequences of distinct behaviors like binge/purge behavior (Smyth et al., 2007) or self-injuring behavior (Muehlenkamp et al., 2009). Interactive sampling approaches can be implemented only on electronic devices (see section "Hardware and Software Solutions"). As mentioned, interactivity in assessment and branching can be used to reduce patients' burden by avoiding the administration of irrelevant questions (e.g., using a conditional question that appears only when tension is significantly increased or decreased; Stiglmayr et al., 2005).

Implications/recommendations regarding time-based designs. In general, time-based sampling schemes can vary in sampling rate (e.g., once daily, five times daily, every hour during waking hours) and timing (i.e., fixed or random intervals) as well as in length of the sampling episode (e.g., 24 hours, one week, or one month). The most important annotation regarding time-based designs is that the sampling design must fit the temporal dynamics of the target processes (Ebner-Priemer & Sawitzki, 2007). Therefore, the temporal resolution of the data is determined by the frequency of repeated assessments. Specifying a time-based sampling protocol requires clarity about the theoretical considerations underlying dynamics in symptomatology, as well as considerations of the expected rapidity of changes in the phenomenon of interest. For example, because BPD is defined by rapid shifts in mood observed over several hours within a day, it may not be informative to use once-a-day diary entries as they cannot capture repeated affective changes over the course of the day. On the other hand, a high frequency sampling with momentary mood ratings every minute in patients with a bipolar disorder is too high, as cycling between manic and depressive episodes is considered to be much slower, with episodes lasting three to six months. This example shows the far-reaching consequences if the sampling design and the process of interest do not match. If the sampling rate is too low, and thus intervals between assessments are too long, the design may fail to uncover natural cycles, may exclude important processes, or may foster biased retrospection. In contrast, if intervals are too short, participants' burden may increase, endangering their compliance, without offering any incremental information over a lower sampling rate.

Despite these important considerations regarding the timing of assessments, there is a paucity of studies comparing various time-based designs, and at present there are no general conventions. This is not surprising, as the temporal dynamics of emotional and cognitive processes are largely unknown and vary across studied phenomena. Ebner-Priemer and Sawitzki (2007) proposed several approaches to investigate the appropriateness of a selected time-based design for the investigation of affective instability. The authors reported various graphical strategies and statistical comparisons that can help to determine whether a specific process has been captured with the chosen time-based design. Future research is needed to go beyond reliance on broad heuristics for deciding on a time-based sampling design and use recommendations from ESM studies addressing the temporal dynamics of the specific affective and cognitive processes being investigated. Furthermore, researchers should provide a clear rationale for their choice of sampling design. Importantly, ESM studies designed with a time-based assessment protocol necessitate the use of some kind of an electronic device to prompt participants whenever an assessment is required. An alert function is especially important when a random time-sampling protocol is used (i.e., repeated assessments at randomly selected times with a predefined minimum and maximum interval between prompts). In our opinion, the best way to do this is to use e-diaries (e.g., Ebner-Priemer et al., 2007b, 2007c; Trull et al., 2008). Although a variety of studies have used paper-and-pencil-diaries in combination with a programmed beeper (e.g., Links et al., 2007) or digital wristwatch (e.g., Myin-Germeys et al., 2001, 2003), the results of a seminal paper investigating the enhancement in compliance achieved with signaling participants with a programmed wristwatch calls the accuracy of paper recordings into question (Broderick, Schwartz, Shiffman, Hufford, & Stone, 2003). We will discuss this issue later in the section "Hardware and Software Solutions."

Implications/recommendations regarding event-based designs. When the target events are rare or occur in an irregular, random manner, event-based designs can be very helpful. Because data (i.e., experiences, attitudes, or behaviors) are gathered only around the predefined event, participants' burden is kept to a minimum and the possible problem of biased retrospection (when events are very rare and

thus the time interval between actual occurrence and time-based assessment may be large) is avoided. The typical procedure used in event-based sampling is to have participants self-initiate an assessment when an event occurs. Therefore, clearly instructing participants on what constitutes the event of interest is important. Another issue in event-based assessment is compliance. Event-based designs do not allow for verifying compliance with the protocol. If the inquiries following a report of an event are burdensome, participants may skip reporting events to avoid completing an assessment. Unfortunately, it is almost impossible to detect such instances of noncompliance. This lack of control constitutes the great disadvantage of event-based sampling schedules.

Even though no general recommendations regarding the length of the assessment period can be offered, two considerations must be carefully balanced: the assessment period has to be (a) sufficiently long to avoid the risk of assessing only a few events of interest but (b) short enough to keep participants on board and avoid compliance problems.

Implications/recommendations regarding combined sampling protocols. Because combined sampling protocols integrate time- and event-contingent sampling, recommendations for both approaches must be taken into account. The expected frequency of the target event should be used to determine the length of the assessment period (to warrant the assessment of a reasonable number of events). Furthermore, the time-based assessment independent of the event should be designed to be as short an interval as necessary to obtain an ample temporal resolution. However, special attention has to be paid to participants' burden, because only compliance with the time-based schedule can be verified. Therefore, the extent of the inquiries should be kept as short as possible.

Implications/recommendations regarding physiology-triggered assessments. Physiology-triggered assessments constitute a strategy closely related to event-based assessment, because a physiological state (e.g., heart rate, activity level) represents an event that initiates a psychological assessment. However, an important difference between the two sampling strategies is that in a physiology-triggered assessment participants do not have to determine if the event occurred because the physiology recorder initiates the assessment. This has the advantage that researchers can verify participants' compliance. Regardless of the wide range of possible applications, we are aware of only two research groups using a physiology-triggered assessment protocol (Kanning, 2011; Myrtek, 2004).

Participants' Acceptance, Burden, Compliance, and Reactivity

The issues of participants' acceptance of ESM, the perceived burden, compliance rates, and methodological reactivity are all highly related. Keeping subjects' perceived burden as low as possible is essential to ensure high compliance and to avoid reactivity. In our own studies, we generally interview participants after the ESM assessment period, asking them about assessment burden and about altered experience or behavior due to assessments (i.e., whether they experienced higher attention to emotions or to bodily sensations). Our very positive experiences are in line with reports of other researchers, because the acceptance of ESM in general seems to be high (Fahrenberg et al., 2007). The same applies to participants' compliance, which is usually very good (Ebner-Priemer & Trull, 2009b; Heron & Smyth, 2010; Hufford, 2007; Mehl & Holleran, 2007).

Various applications of e-diaries in participants drawn from a variety of clinical populations report compliance with the sampling protocol to be as high as 85 percent or higher (e.g., Collins et al., 1998; Ebner-Priemer et al., 2007b, 2007c; Trull et al., 2008). Unfortunately, compliance is hard to prove in event-contingent and combined sampling protocols; however, we think that if the frequency, length, and timing of assessments and the length of the assessment period itself are carefully balanced, high compliance rates can be achieved. Furthermore, certain study procedures (drawn from our experiences across multiple studies we have carried out) that enhance patients' adherence to the assessment protocol include (a) good training, (b) offering to provide feedback about the personal data after the assessment, and (c) graduated study compensation depending on the number of completed data entries.

Concerning reactivity, Fahrenberg and colleagues (2007) pointed out that methodological reactivity is not a specific feature of ESM but rather a property of assessment methods in general. Fortunately, studies have not found much evidence for the influence of assessment methods itself on ESM reports (e.g., Cruise, Broderick, Porter, Kaell, & Stone, 1996; Ebner-Priemer et al., 2007b, 2007c). Additionally, Heron and Smyth (2010) noted that reactivity may be higher when constructs are assessed of which participants are not commonly aware, compared to

more salient experiences or processes (e.g., emotional states, states of distress, pain levels). A recent, interesting investigation of reactivity within ESM was conducted by Clemes, Matchett, and Wane (2008). The authors used pedometers to assess physical activity and found that merely being aware of wearing a pedometer directly increased physical activity. In a follow-up study, Clemes and Parker (2009) disentangled the causes of reactivity, showing that instructing participants to log their step counts in an activity diary daily had the strongest effect in increasing activity relative to a condition in which participants were unaware that they were carrying a pedometer. Aside from the reactivity issue, this study indicates that providing immediate feedback is a powerful feature with regards to motivational aspects, which is in line with the considerations concerning ESM interventions.

That people are interested in tracking their behavior is evidenced by developments in the consumer market. Pedometers and heart rate monitors are sold in sports and discount stores, and telecommunication companies offer software applications that enable mobile phones to measure and track daily steps (e.g., Nokia StepCounter), mood states (e.g., Moody Me; http://itunes.apple.com/us/app/moody-me-mood-diary-tracker/id411567371?mt=8), food intake and calorie consumption (My Diet Diary; http://itunes.apple.com/us/app/my-diet-diary-calorie-counter/id414169919?mt=8), or even Parkinsonian tremor (Lemoyne, Mastroianni, Cozza, Coroian, & Grundfest, 2010). An extensive list of examples can be found in an easy-to-read article in the *New York Times Magazine* (http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html?_r=1&pagewanted=all). Even though these examples all demonstrate the interest of people in tracking their experiences and behaviors, there do remain some privacy concerns about such software applications.

In sum, acceptance, compliance, and reactivity clearly depend on the sampling strategy: the number and frequency of the assessment points, the length of each inquiry, and the length of the entire assessment period. Increasing one or more of these factors will likely increase the assessment burden, and this may reduce acceptance, compliance, and reactivity. Therefore, it is very important to keep both the number of repeated inquiries as well as the assessment period as short as possible and the assessments only as long and as frequent as necessary. Furthermore, it is important to assess adherence to the assessment protocol (i.e., compliance of the participants).

Hardware and Software Solutions

ESM can be conducted using a wide variety of media. In general, there are two main categories: (a) diaries, either e-diaries or paper-and-pencil diaries and (b) physiological ambulatory monitoring devices. Diaries are generally suited for assessing self-reports of experiences (e.g., current mood state), attitudes (e.g., regarding romantic partners), or behaviors (e.g., instances of drug abuse). Furthermore, diaries are used to assess momentary context specificities (e.g., where the participant is [at home, out of home] and with whom [alone or with others]). In addition, providing participants with mobile high-performance physiological ambulatory monitoring sensors makes it possible to assess a large variety of physiological and behavioral data (e.g., posture, activity, or breathing patterns, to name just a few).

DIARIES

Computerization represents a major advantage of ESM protocols over traditional diary techniques. In the past, investigators have often used paper-and-pencil diaries, asking participants to complete one or more diary entries per day. The main limitation of this approach is that investigators cannot be sure that the ratings were actually completed at the scheduled times as specified in the research design. Participants may neglect making scheduled ratings and "back-fill" their diaries instead (an instance of noncompliance, as diaries are not completed as scheduled but are filled out immediately prior to encountering the researcher). Stone and colleagues (2002) investigated the phenomenon of back-filling. They used light-sensitive sensors in diary booklets to confirm compliance with the sampling protocol in a sample of 40 chronic pain patients (another 40 patients were assigned to an e-diary group). In this methodological study, only 11 percent of paper diary reports were answered in accordance with the time schedule (according to the light sensors), although participants reported having completed 90 percent of all reports on time (i.e., the self-reported compliance rate)! This is a significant problem, as low compliance is a serious threat to the validity of the research, even more so when participants fake good compliance by back-filling. Using digital wristwatches that emit auditory signals cannot solve this problem. For example, one study found that self-reported compliance was as high as 85 percent but verified compliance as low as 29 percent (Broderick et al., 2003). The advantage of using e-diaries is that participants complete

ratings in response to prompts emitted by the electronic device. Entries are only allowed within a certain time window after the prompt and are, in addition, electronically time-stamped. Both features circumvent back-filling. Researchers can use these time stamps to verify compliance against the original sampling protocol and hence determine the exact number of missing (or delayed) self-reports (Stone & Shiffman, 2002).

Thus, e-diaries are often considered the gold standard in examining daily life experiences (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). The mere fact that one knows that one's entries are time-stamped apparently heightens compliance. In the aforementioned study by Stone and colleagues (2002), the pain patients using e-diaries showed an actual compliance rate of 94 percent (in contrast to 11 percent of paper diary reports without auditory signals and 29 percent with auditory signals; Broderick et al., 2003; Stone et al., 2002). However, the debate regarding paper-and-pencil versus e-diary methods remains lively (see Green, Rafaeli, Bolger, Shrout, & Reis, 2006 as well as Bolger, Shrout, Green, Rafaeli, & Reis, 2006; Broderick & Stone, 2006; Takarangi, Garry, & Loftus, 2006; Tennen, Affleck, Coyne, Larsen, & DeLongis, 2006, for a detailed discussion).

For some time, handheld computers (also termed personal digital assistants [PDAs]) have been the most commonly used devices for implementing e-diary assessment in research. However, because electronics is a rapidly evolving field with short product cycles, PDAs are being displaced by smartphones and thus are used less and less. For the programming of e-diaries implemented on PDAs, there are a multitude of both free software solutions as well as commercially distributed products (Ebner-Priemer & Kubiak, 2007; Kubiak & Krog, 2011; see also www.ambulatory-assessment.org, which provides links to various commercial and noncommercial software providers). Many software options are also available for programming smartphones, such as the flexible and convenient open source mobile data collection software MyExperience, which can be used for both PDAs and mobile phones with Windows Mobile operating system (available under http://myexperience.sourceforge.net). In addition, there are software solutions for other popular electronic devices like the Apple iPod touch, Sony PlayStation Portable, and Nintendo DS (Kubiak & Krog, 2011).

Independent of the device, researchers planning an ESM study should carefully consider the capabilities as well as limitations of available software solutions. The criteria for choosing a software solution should be mainly based on the features needed for answering a specific research question. Kubiak and Krog (2011) as well as Ebner-Priemer and Kubiak (2007) offer an excellent overview of the functions and features that software should provide. For example, the software should allow for a variety of item formats (e.g., item list, checkboxes, drop-down list, radio buttons, visual analogue scales, and text fields). Furthermore, the software should support various sampling schemes, including time- and event-contingent sampling schemes, as well as their combination. Randomized time sampling (e.g., 8 times between 10 a.m. and 10 p.m. with a predefined minimum interval of 60 minutes between assessments), sampling according to a fixed time schedule (e.g., at 8 a.m., noon, 4 p.m., and 8 p.m.), and the option of defining distinct events associated with different question sets are also desirable features. The software should offer an appropriate signaling function with sound, vibration, visual signals, or a combination of all three, and, most importantly, should time-stamp entries. Concerning signaling, the availability of a snooze or "do not disturb" function that allows participants to temporarily suppress signaling or at least switch into silent mode with vibration and/or visual signals is an important feature to increase adherence to the sampling scheme. The ability to manage assessment logic, including branching within question sets, can be very useful. To use multimodal assessments, the software should provide external trigger capabilities (as long as the hardware provides an interface allowing for a combined assessment of psychological and physiological data). Various physiological signals (e.g., increases in heart rate) or situational context specificities (e.g., noise or temperature) could be used as external triggers to prompt participants. In addition, software solutions for mobile phones support functions that allow participants to record their responses audibly and take pictures or videos, opening ESM for a variety of new application areas and research objects. Even though the initial costs of mobile devices are higher than traditional questionnaire approaches, suitable mobile devices are becoming cheaper and offer the advantage of providing data in electronic format. The transfer of data from paper into electronic format is a task that can be both error-prone and expensive. Finally, however, technological solutions have the disadvantage of higher starting costs, which may limit investigators' ability to conduct pilot studies.

PHYSIOLOGICAL AMBULATORY MONITORING SOLUTIONS

Three types of devices can be distinguished: (a) single-channel devices, (b) specialized multichannel devices, and (c) multipurpose multichannel devices.

S*ingle-channel devices* measure only one physiological parameter (e.g., watches that monitor heart rate). They usually have a preset configuration and are quite simple to handle with regards to assessment and visual data preparation. *Specialized multichannel devices* can measure more than one physiological parameter. Most of these devices have a preset configuration (i.e., many features, like the physiological parameters or the selection of channels, are predetermined). Although confining the range of potential application areas, these devices are still relatively simple to use. In contrast, *multipurpose multichannel devices* that use modular, multipurpose designs are much more flexible, supporting the registration of all kinds of analogue signals. However, greater flexibility comes at the price of higher complexity and potential problems with the application software (e.g., if a specific software solution has not been validated for a biosignal of interest, it needs to be developed under one's own initiative—an expensive and time-consuming process).

Just a few years ago, because of limited memory, recording of raw data was impossible and thus researchers could not subsequently inspect the data (e.g., to determine if implausible values could be explained by noise). Today, the miniaturization of memory allows for raw signals to be recorded at a high resolution. A valuable feature of some recording systems is the potential for online analyses, allowing for immediate feedback or physiology-triggered e-diary assessments (i.e., interactive monitoring, as described above). The sticking point is that the physiology-triggered e-diary assessment is not a standard feature in such recording systems, and it usually has to be programmed by the researcher. However, some software providers are working on the implementation of this feature (see Kubiak & Krog, 2011). Once again, we want to draw attention to the www.ambulatory-assessment.org website because—apart from listing software solutions—it offers an extensive overview of hardware solutions.

Statistical and Analytic Issues

Data obtained by ESM have a complex structure (Bolger, Davis, & Rafaeli, 2003; Schwartz & Stone, 2007). Repeated within-person assessments are obtained, and these multiple assessments of the same person cannot be assumed to be independent. Thus, ESM data show a hierarchical structure (i.e., multiple assessment points are nested within subjects, resulting in the distinction between hierarchical levels). Generally, in ESM data a distinction is drawn between level 1, *momentary* level (e.g., current mood, interpersonal contact), and level 2, *person* level (e.g., age, sex, personality, comorbidity), data. Even though compliance is generally high in ESM studies, instances of missing data are almost always present due to the sheer number of assessments. Thus, the number of repeated assessments is large but not the same for all participants (due to unexpected missing data points). Furthermore, repeated observations of one person often show a serial dependency (i.e., assessments closer in time may be more similar than assessments separated more in time). In addition, neither random sampling nor event sampling nor combined sampling strategies provide data points that are equally spaced in time. Finally, ESM datasets often show temporal patterns and cycles. These complexities in ESM data require the use of flexible mathematical models, since the underlying assumptions of more traditional analytic methods such as the repeated-measures ANOVA are rarely met. Therefore, multilevel models (also called multilinear models, hierarchical linear models, general mixed models, and random regression models; Bolger et al., 2003) are the method of choice and the primary tool for analyses of ESM data.

Various major textbooks offer comprehensive guidelines for building and testing multilevel models. Examples include Hox (2010), Raudenbush and Bryk (2002), as well as Snijders and Bosker (2011). Furthermore, Singer and Willett (2003) discuss longitudinal analysis with a special emphasis on multilevel modeling of individual change. Their book is especially well suited for beginners because it emphasizes data analysis (and not primarily underlying theory) and offers sample codes and data files in a variety of software packages (including SAS, Stata, SPSS, MLwiN, and HLM), allowing the reader to replicate the analyses described in the book. These sample analyses cover a wide range of research questions. In addition, several book chapters dedicated to multilevel modeling using ESM data (Nezlek, 2011; Schwartz & Stone, 2007) as well as papers on specific aspects of ESM data analyses (Jahng, Wood, & Trull, 2008; Kubiak & Jonas, 2007; Nezlek, 2001, 2008) have been published. Approaches to calculate reliability, validity, and sensitivity of change for multilevel data are also available (Wilhelm & Schoebi, 2007). The reader

is also referred to chapter 16 for consideration of latent growth curve modeling in clinical psychology research.

Although complex, multilevel modeling is the recommended method for ESM data analyses. Here, we list several ways such modeling can be used to address important research questions often posed by ESM investigators (Bolger et al., 2003):

1. ESM data can be used to characterize individual differences (i.e., to investigate an individual's typical or average experience, and to examine differences in these averages between subjects). The underlying idea is that aggregating repeated assessments of individuals should result in a more reliable estimate of average experiences compared to single-point-in-time assessments.
2. As ESM provides longitudinal data with higher (temporal) resolution, data obtained by ESM are often used to investigate within-person changes over time (i.e., temporal sequences of a person's experiences), as well as interindividual differences in these temporal sequences. In contrast to aggregating repeated assessments over time, examining "epochs" of experience requires a consideration of the temporal order (and patterns) of data points.
3. Modeling within-person processes is of particular interest because it offers insight into processes underlying changes in an individual's experiences and allows for investigating interindividual differences in these processes. Thus, this analysis can be used to generate explanatory models of the factors affecting within-person changes.

Graphical Display of ESM Data

Although data presentation depends on the research question, we discuss several general approaches to ESM data presentation.

When investigating the antecedents and consequences of certain events, it may be advisable to choose an event-centered data presentation in which the data depicted show multiple assessment points before and after the event of interest (e.g., the event-centered trends of affective antecedents and consequences of self-injuries in patients with bulimia nervosa; Muehlenkamp et al., 2009, p. 86). Another possible approach is the presentation of the case study, as in Figure 11.2 (Santangelo et al., 2010). This approach is useful for providing an overview of the dynamics and the interaction between psychological variables and specific events. Because ESM data provide multiple assessments in temporal order, frequently asked research questions concern the course of a variable of interest over a certain period of time.

For presentation purposes, the individual courses of all participants are often averaged and only group-level patterns are plotted in figures. However, averaging scores within group does not take advantage of the dimensions of ESM data—the time, the subject, and the respective values of the variables of interest. Fortunately, there are several promising examples for graphically visualizing ESM data that do account for the complexity of the data. One example is the graphical description of affective instability (Ebner-Priemer & Sawitzki, 2007). In this study, the investigators used e-diaries to repeatedly assess momentary affective states and states of distress every 10 to 20 minutes over the course of a day in 50 patients with BPD and 50 HCs, resulting in approximately 50 data points per subject. The investigators used R (http://www.r-project.org/) to plot instability of distress ratings for patients and HCs in three dimensions (Fig. 11.3). In this figure, each line (*y*-axis) represents a participant (lines 1 to 50 represent the BPD patients and lines 51 to 100 the HCs), each square represents one of the approximately 50 data points per participant, and the varying shades of gray signify the level of distress (dark shades signify high distress, bright shades signify low distress). It can be easily recognized that the upper half of the figure represents low distress in the group of HCs, with squares mostly colored in bright shades (signifying low distress). In contrast, the lower half of the figure depicts squares mostly colored in dark shades, representing medium and high distress ratings of the BPD patients. The frequently and fast-changing shades of gray in the lower half of the figure represent the well-known affective instability in the BPD group. Kuppens, Oravecz, and Tuerlinckx (2010) offer other examples for the visualization of single time courses in two-dimensional affective space by using affect trajectories and pooled vector field plots. We also direct the reader to the homepage of the R graph gallery (http://addictedtor.free.fr/graphiques/), which offers a large variety of data presentation possibilities with free source codes.

Conclusion

ESM is being used to address research questions in a multitude of mental health disorders, providing new insights into patients' symptoms as they unfold in everyday life. However, contemporary assessment in clinical psychology is still dominated by

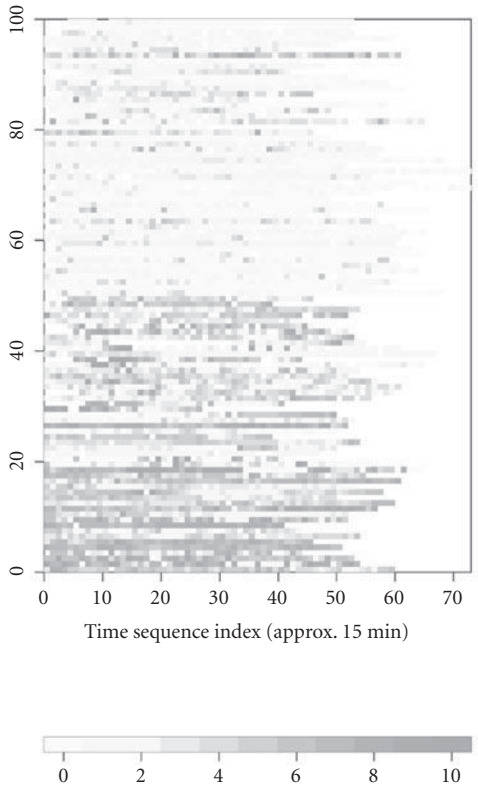


Figure 11.3 This three-dimensional plot shows instability of distress ratings in a group of BPD patients (lines 1 to 50) and HCs (lines 51 to 100) covering the subject, the time, and the respective values of the variables of interest. Each line represents a subject, each square a self-report (with an interval of approximately 15 minutes), and the varying shades of gray the level of distress (bright shades represent self-reports of low distress, dark shades self-reports of high distress). The frequently and fast changing colors in the lower half of the figure represent the affective instability characteristic of patients with BPD.

interview and questionnaire methods. ESM offers several advantages over these traditional assessment methods: (a) real-time assessment, (b) assessment in real-life situations, (c) repeated assessment, (d) multimodal assessment, (e) the possibility of assessing the context of the report, and (f) the possibility of giving feedback in real time. In this chapter, we described several studies that demonstrate the feasibility and utility of ESM in clinical psychology research.

As noted, the main issue in planning and designing an ESM study is that the research question must guide every decision. That means that all of the following are dictated by the research question: (a) the adequacy of the sampling strategy (i.e., time-contingent, event-contingent, or combined sampling), which in turn affects (b) participants' burden, compliance, and reactivity and (c) the choice of hardware and software, (d) the data analytic strategy, and (e) the graphical data description. With regard to the adequacy of the sampling strategy, the sampling design must fit the processes of interest; otherwise, findings may be misleading. Participants' acceptance, the perceived burdensomeness, compliance rates, and methodological reactivity are tightly interwoven. However, many of these factors endangering the significance of findings can be addressed through good patient training, the provision of personal feedback after the assessment, or graduated payments depending on the percentage of completed data entries.

When planning an ESM study, one must decide whether to collect psychological data (using diary methods) or physiological data (using physiological ambulatory monitoring devices) or both. Regarding diary methods, the major decision is to choose either electronic or paper-and-pencil versions. Even though novices to ESM may be discouraged by the substantial initial costs of the hardware, we think that the investment in equipment (i.e., e-diaries or physiological ambulatory monitoring devices) is worth it. Fortunately, hardware used for ESM is becoming increasingly affordable. In addition, the ready availability of both commercial and noncommercial software packages allows for the programming of equipment to pursue a multitude of research questions. Taken together, even though ESM is sometimes seen as more burdensome than traditional questionnaire/interview research, for many research questions it is worth the burden as it provides detailed insight into patients' everyday lives. Furthermore, the possibility of giving real-time support in patients' everyday lives expands ESM beyond a pure assessment strategy and constitutes a promising augmentation to standard treatments.

Future Directions

Sampling Strategies

As previously mentioned, there is a great need for empirically based recommendations for time-based sampling designs. We hope that in the near future more ESM studies address the temporal dynamics of affective and cognitive processes, enabling empirical-based recommendations for time-based designs.

Added Value of ESM Studies

Even though it is tempting to assume that assessing symptoms in everyday life (i.e., where they

actually occur) enhances validity, the incremental, predictive value of ESM data needs to be demonstrated empirically. To this point in time, ESM studies documenting greater ecological and external validity compared to laboratory or questionnaire/interview studies are relatively rare (Ebner-Priemer & Trull, 2009a). However, we are convinced that ESM provides additional information and reveals a more complete picture of the symptomatology. Proceeding from the assumption that ESM data are more valid for many study questions, ESM should lead to better predictions compared to retrospective questionnaires and interviews for the study of many phenomena. This assumption is supported by some recent empirical studies (e.g., Stepp, Hallquist, Morse, & Pilkonis, 2010). However, replication is warranted, and thus the empirical investigation of the additional predictive value of ESM data represents a necessary area for future research.

Use of ESM in Clinical Psychology

We discussed the key features of ESM approaches in the section "Characteristics and Advantages of ESM." In our opinion, these key features are compelling. For example, from a behavioral self-management perspective, self-monitoring is very important when trying to change behavior. Therefore, the continuous assessment of one's own experiences, attitudes, or behaviors is a first step in the direction of intended change and self-management. Self-monitoring can be expanded to integrate with real-time feedback in daily life; ESM can be directly instrumentalized as a therapeutic tool, constituting the interface between clinical assessment and clinical treatment. However, certain important questions need to be answered before these techniques can be incorporated into routine practice. First, it must be clarified if the cost/benefit ratio is advantageous (i.e., if comparable outcomes are obtained while the clinician's time and effort are saved). Second, a determination of which patients are open to this form of assessment and treatment is needed. It may be that a continuous tracking of symptoms and thus constant surveillance is not suitable for all patient groups, such as psychotic patients struggling with paranoid ideation (although several examples with patients suffering from severe mental illness are provided by Oorschot et al., 2009). Furthermore, patients' compliance with the electronically prescribed coaching, advice, and practices has to be examined carefully, because patients are asked to use specific and perhaps (to them) novel skills to alleviate their symptoms. This is likely to prove more challenging than simply reporting experience of symptomatology (Solzbacher et al., 2007).

Future Prospects: Gaining a Holistic Picture of Patients' Symptomatology

Modern ESM devices are suitable for collecting not only self-reports but also audio, video, and geographic positioning, as well as physiological and biological data. In recent years the integration of information from wearable devices and biosensors with the information gathered via PDAs and smartphones has become more and more feasible. This development enables researchers and clinicians to gather a comprehensive picture of patients' emotional, psychological, behavioral, and physical functioning in their natural environment. Thus, ESM data can provide a detailed account and understanding of an individual's problems as experienced in daily life. In turn, this information can both inform and enhance clinical treatment.

References

Adamo, K. B., Prince, S. A., Tricco, A. C., Connor-Gorber, S., & Tremblay, M. (2009). A comparison of indirect versus direct measures for assessing physical activity in the pediatric population: A systematic review. *International Journal of Pediatric Obesity*, *4*, 2–27.

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders, third edition (DSM-III)*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders, fourth edition, text revision (DSM-IV-TR)*. Washington, DC: American Psychiatric Association.

Bauer, M., Grof, P., Rasgon, N., Bschor, T., Glenn, T., & Whybrow, P. C. (2006). Temporal relation between sleep and mood in patients with bipolar disorder. *Bipolar Disorders*, *8*, 160–167.

Ben Zeev, D., & Young, M. A. (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: An experience sampling study. *Journal of Nervous and Mental Disease*, *198*, 280–285.

Ben Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition & Emotion*, *23*, 1021–1040.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.

Bolger, N., Shrout, P. E., Green, A. S., Rafaeli, E., & Reis, H. T. (2006). Paper or plastic revisited: Let's keep them both—Reply to Broderick and Stone (2006); Tennen, Affleck, Coyne, Larsen, and DeLongis (2006); and Takarangi, Garry, and Loftus (2006). *Psychological Methods*, *11*, 123–125.

Broderick, J. E., Schwartz, J. E., Shiffman, S. S., Hufford, M. R., & Stone, A. A. (2003). Signaling does not adequately improve diary compliance. *Annals of Behavioral Medicine*, *26*, 139–148.

Broderick, J. E., & Stone, A. A. (2006). Paper and electronic diaries: Too early for conclusions on compliance rates and

their effects—Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). *Psychological Methods*, *11*, 106–111.

Bussmann, H. B. J., Ebner-Priemer, U. W., & Fahrenberg, J. (2009). Ambulatory behavior monitoring: Progress in measurement of activity, posture, and specific motion patterns in daily life. *European Psychologist*, *14*, 142–152.

Chambless, D. L., Caputo, G. C., Jasin, S. E., Gracely, E. J., & Williams, C. (1985). The mobility inventory for agoraphobia. *Behaviour Research and Therapy*, *23*, 35–44.

Clemes, S. A., Matchett, N., & Wane, S. L. (2008). Reactivity: An issue for short-term pedometer studies? *British Journal of Sports Medicine*, *42*, 68–70.

Clemes, S. A., & Parker, R. A. (2009). Increasing our understanding of reactivity to pedometers in adults. *Medicine and Science in Sports and Exercise*, *41*, 675–681

Collins, R. L., Morsheimer, E. T., Shiffman, S. S., Paty, J. A., Gnys, M., & Papandonatos, G. D. (1998). Ecological momentary assessment in a behavioral drinking moderation training program. *Experimental and Clinical Psychopharmacology*, *6*, 306–315.

Conner, T., & Barrett, L. F. (2012). Trends in ambulatory self-report: Understanding the utility of momentary experiences, memories, and beliefs. *Psychosomatic Medicine, 74,* 327–337.

Cruise, C. E., Broderick, J. E., Porter, L. S., Kaell, A., & Stone, A. A. (1996). Reactive effects of diary self-assessment in chronic pain patients. *Pain*, *67*, 253–258.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*, 526–536.

de Beurs, E., Lange, A., & van Dyck, R. (1992). Self-monitoring of panic attacks and retrospective estimates of panic: Discordant findings. *Behaviour Research & Therapy*, *30*, 411–413.

Delespaul, P. A., & de Vries, M. W. (1987). The daily life of ambulatory chronic mental patients. *Journal of Nervous and Mental Disease*, *175*, 537–544.

Ebner-Priemer, U. W., Bohus, M., & Kuo, J. (2007a). Can retrospective interviews assess instability? A comparison of ambulatory assessment and expert interviews regarding DSM-IV criteria for borderline personality disorder. In M. J. Sorbi, H.Rüddel, & M.Bühring (Eds.), *Frontiers in stepped ecare* (pp. 25–37). Utrecht: University Utrecht.

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T.J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology*, *188*, 195–202.

Ebner-Priemer, U. W., & Kubiak, T. (2007). Psychological and psychophysiological ambulatory monitoring—A review of hardware and software solutions. *European Journal of Psychological Assessment*, *23*, 214–226.

Ebner-Priemer, U. W., Kuo, J., Kleindienst, N., Welch, S. S., Reisch, T., Reinhard, I., et al. (2007b). State affective instability in borderline personality disorder assessed by ambulatory monitoring. *Psychological Medicine*, *37*, 961–970.

Ebner-Priemer, U. W., Kuo, J., Welch, S. S., Thielgen, T., Witte, S., Bohus, M., et al. (2006). A valence-dependent group-specific recall bias of retrospective self-reports: A study of borderline personality disorder in everyday life. *Journal of Nervous and Mental Disease*, *194*, 774–779.

Ebner-Priemer, U. W., & Sawitzki, G. (2007). Ambulatory assessment of affective instability in borderline personality disorder—The effect of the sampling frequency. *European Journal of Psychological Assessment*, *23*, 238–247.

Ebner-Priemer, U. W., & Trull, T. J. (2009a). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment*, *21*, 463–475.

Ebner-Priemer, U. W., & Trull, T. J. (2009b). Ambulatory assessment—An innovative and promising approach for clinical psychology. *European Psychologist*, *14*, 109–119.

Ebner-Priemer, U. W., Welch, S. S., Grossman, P., Reisch, T., Linehan, M. M., & Bohus, M. (2007c). Psychophysiological ambulatory assessment of affective dysregulation in borderline personality disorder. *Psychiatry Research*, *150*, 265–275.

Fahrenberg, J. (1996). Ambulatory assessment: Issues and perspectives. In J.Fahrenberg & M.Myrtek (Eds.), *Ambulatory assessment. Computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 3–20). Seattle: Hogrefe & Huber.

Fahrenberg, J., Hüttner, P., & Leonhart, R. (2001). MONITOR: Acquisition of psychological data by a hand-held PC. In J. Fahrenberg & M. Myrtek (Eds.), *Progress in ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies* (pp. 93–112). Seattle: Hogrefe & Huber.

Fahrenberg, J., & Myrtek, M. (1996). *Ambulatory assessment. Computer-assisted psychological and psychophysiological methods in monitoring and field studies*. Seattle: Hogrefe & Huber.

Fahrenberg, J., & Myrtek, M. (2001). Ambulantes monitoring und assessment. In F. Rösler (Ed.), *Grundlagen und Methoden der Psychophysiologie* (pp. 657–796). Göttingen: Hogrefe.

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment—Monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, *23*, 206–213.

FDA—U. S. Food and Drug Administration (2009). *Guidance for industry—Patient-reported outcome measures: Use in medical product development to support labeling claims*. Retrieved May 1, 2011, from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, *14*, 577–606.

Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, *11*, 87–105.

Hansen, T. W., Jeppesen, J. R., Rasmussen, S., Ibsen, H., & Torp-Pedersen, C. (2006). Ambulatory blood pressure monitoring and risk of cardiovascular disease: A population-based study. *American Journal of Hypertension*, *19*, 243–250.

Harvey, P. D., Greenberg, B. R., & Serper, M. R. (1989). The Affective Lability Scales: Development, reliability, and validity. *Journal of Clinical Psychology*, *45*, 786–793.

Herman, S., & Koran, L. M. (1998). In vivo measurement of obsessive-compulsive disorder symptoms using palmtop computers. *Computers in Human Behavior*, *14*, 449–462.

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*, 1–39.

Horemans, H. L., Bussmann, J. B., Beelen, A., Stam, H. J., & Nollet, F. (2005). Walking in postpoliomyelitis syndrome: The relationships between time-scored tests, walking in daily life and perceived mobility problems. *Journal of Rehabilitation Medicine*, *37*, 142–146.

Houtveen, J. H., & de Geus, E. J. (2009). Noninvasive psychophysiological ambulatory recordings: Study design and data analysis strategies. *European Psychologist*, *4*, 132–141.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge Academic.

Hufford, M. R. (2007). Special methodological challenges and opportunities in ecological momentary assessment. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 54–75). Oxford: University Press.

Intille, S. S. (2007). Technological innovations enabling automatic, context-sensitive ecological momentary assessment. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 308–337). Oxford: University Press.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*, 354–375.

Kahneman, D., Fredrickson, B. L., Schreiber, C., & Redelmeier, D. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*, 401–405.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*, 1776–1780.

Kanning, M. (2011, January). *The effect of physical activity in everyday life on different mood states using an interactive ambulatory assessment*. Talk presented at the Symposium for Ambulatory Assessment, Karlsruhe Institute of Technology, January 12, 2011, Karlsruhe, Germany.

Kenardy, J. A., Dow, M. G., Johnston, D. W., Newman, M. G., Thomson, A., & Taylor, C. B. (2003). A comparison of delivery methods of cognitive-behavioral therapy for panic disorder: An international multicenter trial. *Journal of Consulting and Clinical Psychology*, *71*, 1068–1075.

Kihlstrom, J. F., Eich, E., Sandbrand, D., & Tobias, B. A. (2000). Emotion and memory: Implications for self-report. In A. A. Stone & J. S. Turkkan (Eds.), *Science of self- report: Implications for research and practice* (pp. 81–99). Mahwah, NJ: Lawrence Erlbaum Associates.

Kubiak, T., & Jonas, C. (2007). Applying circular statistics to the analysis of monitoring data. *European Journal of Psychological Assessment*, *23*, 227–237.

Kubiak, T., & Krog, K. (2011). Computerized sampling of experiences and behavior. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 124 – 143). New York: Guilford.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*, 1042–1060.

Larsen, R. J., Diener, E., & Emmons, R. A. (1986). Affect intensity and reactions to daily life events. *Journal of Personality and Social Psychology*, *51*, 803–814.

Lemke, M. R., Broderick, A., Zeitelberger, M., & Hartmann, W. (1997). Motor activity and daily variation of symptom intensity in depressed patients. *Neuropsychobiology*, *36*, 57–61.

Lemoyne, R., Mastroianni, T., Cozza, M., Coroian, C., & Grundfest, W. (2010). Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless accelerometer application. *Conference Proceedings IEEE Engineering in Medicine and Biology Society 2010*, 4954–4958.

Linehan, M. M. (1993). *Skills training manual for treating borderline personality disorder*. New York: Guilford.

Links, P. S., Eynan, R., Heisel, M. J., Barr, A., Korzekwa, M., McMain, S., et al. (2007). Affective instability and suicidal ideation and behavior in patients with borderline personality disorder. *Journal of Personality Disorder*, *21*, 72–86.

Margraf, J., Taylor, C. B., Ehlers, A., Roth, W. T., & Agras, W. S. (1987). Panic attacks in the natural environment. *Journal of Nervous and Mental Disease*, *175*, 558–565.

Marks, M., & Hemsley, D. (1999). Retrospective versus prospective self-rating of anxiety symptoms and cognitions. *Journal of Anxiety Disorders*, *13*, 463–472.

Mehl, M. R., & Holleran, S. E. (2007). An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (EAR). *European Journal of Psychological Assessment*, *23*, 248–257.

Meuret, A. E., Wilhelm, F. H., Ritz, T., & Roth, W. T. (2008). Feedback of end-tidal pCO2 as a therapeutic approach for panic disorder. *Journal of Psychiatric Research*, *42*, 560–568.

Morey, L. C. (1991). *Personality assessment inventory: Professional manual*. Odessa: Psychological Assessment Resources.

Muehlenkamp, J. J., Engel, S. G., Wadeson, A., Crosby, R. D., Wonderlich, S. A., Simonich, H., et al. (2009). Emotional states preceding and following acts of non-suicidal self-injury in bulimia nervosa patients. *Behaviour Research & Therapy*, *47*, 83–87.

Murray, G. (2007). Diurnal mood variation in depression: A signal of disturbed circadian function? *Journal of Affective Disorders*, *102*, 47–53.

Myin-Germeys, I., Peeters, F., Havermans, R., Nicolson, N. A., de Vries, M. W., Delespaul, P. A., et al. (2003). Emotional reactivity to daily life stress in psychosis and affective disorder: An experience sampling study. *Acta Psychiatrica Scandinavica, Supplementum*, *107*, 124–131.

Myin-Germeys, I., van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. *Archives of General Psychiatry*, *58*, 1137–1144.

Myrtek, M. (2004). *Heart and emotion: Ambulatory monitoring studies in everyday life*. Seattle: Hogrefe & Huber.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*, 771–785.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*, 842–860.

Nezlek, J. B. (2011). Multi-level modeling analysis of diary-style data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 357–383). New York: Guilford.

Oorschot, M., Kwapil, T., Delespaul, P., & Myin-Germeys, I. (2009). Momentary assessment research in psychosis. *Psychological Assessment*, *21*, 498–505.

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion*, *6*, 383–391.

Piasecki, T. M., Hufford, M. R., Solhan, M., & Trull, T. J. (2007). Assessing clients in their natural environments with electronic diaries: Rationale, benefits, limitations, and barriers. *Psychological Assessment*, *19*, 25–43.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus

self-report measures for assessing physical activity in adults: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *5*, 56.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Russell, J. J., Moskowitz, D. S., Zuroff, D. C., Sookman, D., & Paris, J. (2007). Stability and variability of affective experience and interpersonal behavior in borderline personality disorder. *Journal of Abnormal Psychology*, *116*, 578–588.

Salles, G. F., Cardoso, C. R. L., & Muxfeldt, E. S. (2008). Prognostic influence of office and ambulatory blood pressures in resistant hypertension. *Archives of Internal Medicine*, *168*, 2340–2346.

Santangelo, P., Ebner-Priemer, U.W., Koudela, S., & Bohus, M. (2010, July). *Does dysfunctional behavior improve affect and distress in everyday life?* Talk presented at the 1st International Congress on Borderline Personality Disorder, July 1–3, 2010, Berlin, Germany.

Schwartz, J. E., & Stone, A. A. (2007). The analysis of real-time momentary data: A practical guide. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 76–113). Oxford: Oxford University Press.

Shiffman, S. S. (2005). Dynamic influences on smoking relapse process. *Journal of Personality*, *73*, 1715–1748.

Shiffman, S. S. (2007). Designing protocols for ecological momentary assessment. In A. A. Stone, S. S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 27–53). New York: Oxford University Press.

Shiffman, S. S. (2009). Ecological momentary assessment (EMA) in studies of substance use. *Psychological Assessment*, *21*, 486–497.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.*New York: Oxford University Press.

Smyth, J. M., & Heron, K. E. (2011). Health psychology. In M. M. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 570–584). New York: Guilford Press.

Smyth, J. M., Wonderlich, S. A., Heron, K. E., Sliwinski, M. J., Crosby, R. D., Mitchell, J. E., et al. (2007). Daily and momentary mood and stress are associated with binge eating and vomiting in bulimia nervosa patients in the natural environment. *Journal of Consulting and Clinical Psychology*, *75*, 629–638.

Smyth, J. M., Wonderlich, S. A., Sliwinski, M. J., Crosby, R. D., Engel, S. G., Mitchell, J. E., et al. (2009). Ecological momentary assessment of affect, stress, and binge-purge behaviors: Day of week and time of day effects in the natural environment. *International Journal of Eating Disorders*, *42*, 429–436.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Solhan, M. B., Trull, T. J., Wood, P. K., & Jahng, S. (2009). Clinical assessment of affective instability: Comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment*, *21*, 425–436.

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. Bühring (Eds.), *Frontiers in stepped care* (pp. 111–119). Utrecht: University Utrecht.

Spitzer, R. L., & William, J. B. (1983). *Structured Clinical Interview for DSM-III—Upjohn version (SCID-UP)*. New York: New York State Psychiatric Institute, Biometrics Research Department.

Stein, K. F., & Corte, C. M. (2003). Ecologic momentary assessment of eating-disordered behaviors. *International Journal of Eating Disorders*, *34*, 349–360.

Stepp, S. D., Hallquist, M. N., Morse, J. Q., & Pilkonis, P. (2010). Multimethod investigation of interpersonal functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*. Advance online publication. doi: 10.1037/a0020572.

Stiglmayr, C., Gratwohl, T., Linehan, M. M., Fahrenberg, J., & Bohus, M. (2005). Aversive tension in patients with borderline personality disorder: A computer-based controlled field study. *Acta Psychiatrica Scandinavica*, *111*, 372–379.

Stiglmayr, C. E., Ebner-Priemer, U. W., Bretz, J., Behm, R., Mohse, M., Lammers, C. H.. et al. (2008). Dissociative symptoms are positively related to stress in borderline personality disorder. *Acta Psychiatrica Scandinavica*, *117*, 139–147.

Stone, A. A., & Broderick, J. E. (2007). Real-time data collection for pain: Appraisal and current status. *Pain Medicine*, *8Supplement* 3, S85-S93.

Stone, A. A. & Shiffman, S. S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*, 236–243.

Stone, A. A., Shiffman, S. S., Atienza, A. A., & Nebeling, L. (2007). *The science of real-time data capture: Self-reports in health research*. New York: Oxford University Press.

Stone, A. A., Shiffman, S. S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *British Medical Journal*, *324*, 1193–1194.

Takarangi, M. K., Garry, M., & Loftus, E. F. (2006). Dear diary, is plastic better than paper? I can't remember: Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). *Psychological Methods*, *11*, 119–122.

Taylor, C. B., Sheikh, J., Agras, W. S., Roth, W. T., Margraf, J., Ehlers, A., et al. (1986). Ambulatory heart rate changes in patients with panic attacks. *American Journal of Psychiatry*, *143*, 478–482.

Tennen, H., Affleck, G., Coyne, J. C., Larsen, R. J., & DeLongis, A. (2006). Paper and plastic in daily diary research: Comment on Green, Rafaeli, Bolger, Shrout, and Reis (2006). *Psychological Methods*, *11*, 112–118.

Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., et al. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, *117*, 647–661.

Tryon, W. W. (2006). Activity measurement. In H. Michel (Ed.), *Clinician`s handbook of adult behavioral assessment* (pp. 85–120). New York: Academic Press.

Tryon, W. W., Tryon, G. S., Kazlausky, T., Gruen, W., & Swanson, J. M. (2006). Reducing hyperactivity with a feedback actigraph: Initial findings. *Clinical Child Psychology and Psychiatry*, *11*, 607–617.

Verdecchia, P., Schillaci, G., Borgioni, C., Ciucci, A., Gattobigio, R., Zampi, I., et al. (1998). Prognostic value of a new electrocardiographic method for diagnosis of left ventricular

hypertrophy in essential hypertension. *Journal of the American College of Cardiology*, *31*, 383–390.

Volkers, A. C., Tulen, J. H. M., Van Den Broek, W. W., Bruijn, J. A., Passchier, J., & Pepplinkhuizen, L. (2003). Motor activity and autonomic cardiac functioning in major depressive disorder. *Journal of Affective Disorders*, *76*, 23–30.

Wang, P. S., Beck, A. L., Berglund, P., McKenas, D. K., Pronk, N. P., Simon, G. E., et al. (2004). Effects of major depression on moment-in-time work performance. *American Journal of Psychiatry*, *161*, 1885–1891.

Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the positive and negative affect schedule—Expanded Form*. Retrieved February 23, 2008, from University of Iowa, web site: http://www.psychology.uiowa.edu/Faculty/Clark/PANAS-X.pdf.

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychiatry*, *84*, 552–569.

Wilhelm, F. H., & Roth, W. T. (1998). Taking the laboratory to the skies: Ambulatory assessment of self-report, autonomic, and respiratory responses in flying phobia. *Psychophysiology*, *35*, 596–606.

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life—Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment*, *23*, 258–267.

World Health Organisation (1992). *The ICD-10: Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organisation.

PART 3

Analytic Strategies for Clinical Psychology

This page intentionally left blank

CHAPTER

12 Statistical Power: Issues and Proper Applications

Helena Chmura Kraemer

Abstract

Statistical hypothesis testing has recently come under fire, as invalid and inadequately powered tests have become more frequent and the results of statistical hypothesis testing have become harder to interpret. Following a review of the classic definition of a statistical test of a hypothesis, discussion of common ways in which this approach has gone awry is presented. The review concludes with a description of the structure to support valid and powerful statistical testing and the context in which power calculations are properly done.

Key Words: Power, hypothesis testing, *a priori* hypothesis, null hypothesis, types I, II, III errors, significance level, a valid test, an adequately powered test, effect sizes, critical effect size, exploratory studies, pilot studies, meta-analysis, equipoise

Introduction

Even after all these years, the definitive book on statistical power remains that written by Cohen (Cohen, 1977, 1988). In the late 1980s, Cohen encouraged me to write a more didactic book (Kraemer & Thiemann, 1987) on statistical power, to make power considerations in comparing various designs more accessible to behavioral and medical researchers. Shortly before his untimely death in 1998, Cohen commented that we may have done more harm than good. When researchers were simply ignoring statistical power, he suggested, reviewers used their experience and common sense to decide whether evidence would be (proposal) or was (submitted paper) convincing. Now, there appears to be a great deal of "mathematical bludgeoning" (Paulos, 1988) with erroneous power calculations, often citing our books, that obscure good commonsense decisions. As a result, he thought, there were probably more, not fewer, poorly designed studies.

Others entertain similar worries. Some have suggested that, because of such misuse and abuse, statistical hypothesis testing might simply be banned (Hunter, 1997; Krantz, 1999; Nickerson, 2000; Shrout, 1997). Yet the statistical hypothesis-testing approach, properly used, is unique, elegant, and useful; it would be a shame to discard such a valued tool without a serious attempt to resolve its problems.

An explanation for this situation might be illustrated by a baseball analogy. If a potential baseball player were well trained in skills, to bat, field, and run, but was unaware of what the rules of baseball were, or how to interact with the other members of his team to win games, that player would make major mistakes were he sent out to play in a game. In the same way, in the many books and papers written on power (Aberson, 2010; Dattalo, 2008; Murphy, Myors, & Wolach, 2009), in the way in which we teach students, with the emphasis on tables, charts, and computer programs available "to do power," the emphasis has been on the *computations* necessary to consider power, not on the *context* in which such computations should be done. Such computation is a small but essential part of designing a valid and adequately powered research study,

and, in turn, statistical hypothesis testing is a small but essential part of research. Consequently, in this presentation, the primary focus is on the "rules of the (scientific) game" and on "(scientific) teamwork," not on computation.

In what follows, the classic definition of a statistical test of a hypothesis is reviewed, principles that have largely existed since the early 20th century. Following that will be discussion of the common ways in which this approach has gone awry, ending with a summary of the structure to support valid and powerful statistical testing, and the context in which power calculations are properly done. While these principles apply to any statistical significance testing, we here focus on a simple common problem for illustration, a randomized clinical trial in which a representative sample from a population is to be randomly assigned to interventions A and B, with response to treatment assessed by evaluators who are "blinded" to treatment assignment.

Classic Statistical Hypothesis Testing

The A Priori Hypothesis

Every proposal starts with a *hypothesis*, a theory the researchers would like to evaluate, that would advance science in their area. The hypothesis must be *a priori* (i.e., articulated before the data to be used to test that hypothesis are accessed). The *null hypothesis* is the denial of that theory. Thus, if the hypothesis is that A is preferable to B (symbolically $A > B$), the null hypothesis is that A is either equivalent or inferior to B ($A \leq B$), a "one-tailed hypothesis." On the other hand, if the hypothesis is more generally that A is different from B ($A \neq B$), the null hypothesis is that A is identical to B ($A = B$), a "two-tailed hypothesis." These are two very different hypotheses that lead to two different tests. To add yet another version: If the hypothesis is that A is different from B both among men and among women in the population, it must be remembered that it may be that $A > B$ among men and $A = B$ among women. The null hypothesis then is that $A = B$ both among men and among women separately, a different hypothesis, requiring a different design, test, and power calculation. The precise articulation of the specific hypothesis to be tested, and thus the associated null hypothesis, is crucial to statistical hypothesis testing.

Theoretical Rationale and Empirical Justification

There must be both theoretical rationale and empirical justification for thinking that the hypothesis proposed might be true and might advance scientific knowledge. Whether that hypothesis is one-tailed or two-tailed, for example, is determined by such rationale and justification. Such rationale and justification is obtained in critical reading of the relevant scientific literature, in clinical observation, in discussions with expert colleagues, in previous research findings, in secondary analyses of datasets compiled to test other hypotheses, and in performing studies not driven by *a priori* hypotheses, done explicitly for the purpose of exploring new scientific areas (exploratory or hypothesis-generating studies). Multiple such sources serve first as the research background for a strong hypothesis, and then to guide the designs of strong and valid studies to test those hypotheses.

Equipoise

For ethical research (particularly with human subjects), there must be enough background information to motivate proposing the hypothesis, but not enough to be convincing to the researchers of the truth of their hypothesis. This is called *equipoise* (Freedman, 1987). The ethical arguments may be left to others, but the practical implications of these requirements are important. If there is not enough information to motivate the hypothesis, there is also not enough to make the design decisions necessary for a valid and adequately powered statistical test of that hypothesis. Then the risk of a failed study, of wasting time and resources, and of putting an unwarranted burden on the participants in the study, is too high. At the other extreme, if the researchers are already convinced that their hypothesis is true, it is very difficult to maintain the objectivity necessary to design a fair test of the hypothesis, or to execute the study and interpret the results without bias. Researchers have been heard to declare that the results of their own study did not "come out right," often subsequently slanting their reports of that study to hide what "came out wrong." When the researchers conducting the study know/believe what the answer must be, the risk of misleading results is too high. Obviously researchers would prefer that their hypothesis be true, but there must be reasonable doubt as to whether or not the hypothesis is true in proposing and testing that hypothesis.

A Valid Statistical Test

A *statistical test* is a rule proposed *a priori*, stipulating what needs to be observed in the dataset to reject the null hypothesis, and thus to declare support for the researchers' theory.

Thus, in a randomized controlled trial in which the N sampled subjects are to be randomized Np to intervention A, and N(1 – p) (0 < p < 1) to intervention B, in order to compare outcomes, we might propose to compute the mean response to A (M_A) and that to B (M_B), and the pooled within-group standard deviation (s_p), and from these compute the *t* statistic = $(Np(1 - p))^{1/2}(M_A - M_B)/s_p$. Then for a one-tailed hypothesis, we might propose to reject the null hypothesis if the *t* statistic is greater than the 95th percentile of the central *t*-distribution with (N – 2) degrees of freedom, and for the two-tailed hypothesis if the absolute value of the *t* statistic were greater than the 97.5th percentile of that distribution. This test is taught in every elementary statistics course and is usually more briefly stated as "We will do a 5-percent-level two-sample *t* test to compare A with B."

Alternatively, we might propose to use the Mann-Whitney-Wilcoxon test, or we might propose to dichotomize response, declaring a participant with response greater than some specified cut-point a "success" and all others a "failure," in which case we would use some appropriate variation of the 2 × 2 chi-square test. These too are familiar tests readily available to users, and there are many others less standard.

Which one to use? That decision is based on evaluating the performance of each proposed test, specifically the probability of rejecting the null hypothesis when the null hypothesis is true (significance level), and then when the researchers' hypothesis is true (power).

Significance Level

A *significance level* is set *a priori*, usually denoted α. A *valid α-level test* is one where the probability of rejecting the null hypotheses *if it is true* is less than α. This type of error, that of reporting a false-positive result, is called a *type I error*. The significance level is set to protect against an avalanche of false-positive results that would otherwise result. Technically, α can be any level set by the researchers, provided it is set *a priori*. Practically, however, there is scientific consensus that α should be .05 (less often, .01), as appropriate for protecting the integrity of scientific findings. Thus, if researchers want to set α at, say, .06 *a priori*, technically they can do so, provided that is acceptable to their scientific peers (in clinical psychology, .05 is the established cut). What is not legitimate is to set α at .05 *before* the study, and then to change it to .06 when the observed results are not quite good enough! To do this is somewhat like trying to change your bet on a horserace after you see which horse actually won.

If the observed responses in the two groups were normally distributed with equal variances, it has been long known that the *t* tests proposed above are valid α-level tests. However, if the distributions are not normal or the variances unequal, those *t* tests are not necessarily valid α-level tests. We might then prefer to use another test, such as a Mann-Whitney-Wilcoxon or chi-square test, both valid in wider circumstances. After we have eliminated tests not valid for the population, design, or measures, based on what we know about the situation from the background research, that will still leave many possible tests among which to choose, as well as the choice of N and p.

To understand these choices and the consequences of making the wrong choice requires an understanding of statistical power, which in turn is based on an understanding of effect sizes.

Effect Size

The *effect size* is a population parameter that reflects the degree of consistency of what is actually going on in the population with the hypothesis proposed. For any hypothesis, there are multiple forms of effect size. Because rejection of the null hypothesis (i.e., a "statistically significant" result) does not necessarily mean that the result is scientifically important (of clinical or practical significance), it is preferable to choose a form of effect size that fosters consideration of such clinical or practical significance (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999).

When the *t* test is valid, the usual choice of effect size is $\delta = (\mu_A - \mu_B)/\sigma$, where the Greek letters denote the (unknown) population means and the standard deviation common to the two populations, estimated in the sample by Cohen's $d = (M_A - M_B)/s_p$ defined above, the standardized mean difference. To foster interpretability, and comparability across situations when distributions are not normal or variances unequal, one might instead choose some function of δ—for example, success rate difference, $SRD = 2\Phi(\delta/\sqrt{2}) - 1$ (where Φ() is the cumulative standard normal distribution function), or number needed to treat, NNT = 1/SRD (Kraemer & Kupfer, 2006). NNT is particularly relevant to clinical psychologists' evaluations of treatment outcomes (Shearer-Underhill & Marker, 2010). Here, a patient is classified as a "success" if he or she has a response preferable to a randomly selected patient from the other group. SRD is the difference between the rates of success in the A versus B groups. NNT

is the number of A patients one would need to sample to expect to have one more "success" than if the same number of B patients were sampled. Neither SRD nor NNT is tied to normal distribution theory, and thus both can be used in situations whether or not the *t* test is valid. SRD is usually preferred for calculations, whereas NNT is often more informative and interpretable.

For a valid α-level test and a well-chosen form of effect size, the researchers' hypothesis can be translated into a statement about effect sizes (e.g., SRD > 0 for a one-tailed test or SRD ≠ 0 for a two-tailed one) as also can the null hypothesis (e.g., SRD ≤ 0 or SRD = 0).

Power

The *power* of a valid α-level test to detect a deviation from the null hypothesis for different values of SRD is the probability of rejecting the null hypothesis for any SRD consistent with the researchers' hypothesis (e.g., SRD > 0 or SRD ≠ 0). Not rejecting the null hypothesis when the researchers' hypothesis is true is called a *type II error*, a false negative. Figures 12.1 and 12.2 are graphs of the probability of rejecting the null hypothesis when the *t*-test assumptions are satisfied for all values of the SRD, when N = 100, p = 1/2. In Figure 12.1 are the graphs for the 5 percent one- and two-tailed *t* tests for all possible values of SRD; in Figure 12.2 are the corresponding graphs for the one-tailed *t* test and the chi-square tests for three different dichotomizations. (For SRD < 0, the curves continue to decline as for the one-tailed *t* test in Fig. 12.1.)

When the assumptions of the *t* test are valid, it can be seen in Figures 12.1 and 12.2 that these are all valid 5 percent tests. For all these tests, the probability of rejecting the null hypothesis is no greater

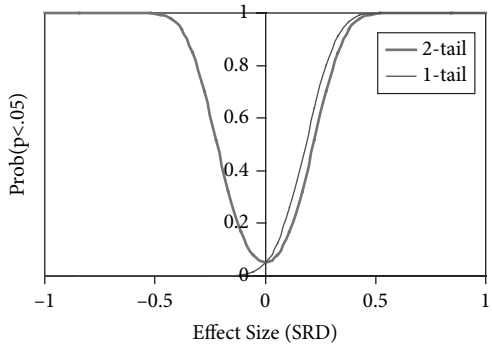


Figure 12.1 The performance curves (probability of a statistically significant result) for valid one- and two-tailed 5-percent-level *t* tests, N = 100 subjects, p = 1/2.

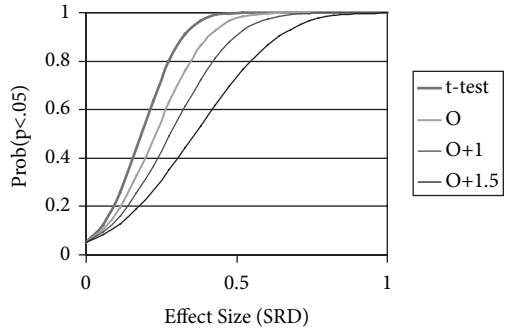


Figure 12.2 The performance curves (probability of a statistically significant result) for a valid one-tailed 5-percent-level *t* tests, N = 100 subjects, p = 1/2, and for three dichotomizations: the optimal one with the cut-point halfway between the two means (O), and the cut-point ±1 (O + 1) and ±1.5 (O + 1.5) standard deviations above or below that point.

than .05 when the null hypothesis is true (SRD ≤ 0 for the one-tailed tests, and SRD = 0 for the two-tailed one). The probability increases as SRD moves away from the null hypothesis values toward stronger and stronger support for the researchers' hypothesis. But, as here, the power for any valid α-level test can lie anywhere between α and 1.0 depending on what the unknown effect size is in the population. What then is an "adequately powered test"?

The Critical Effect Size

We must clearly distinguish now between three different effect sizes, the true population effect size that is unknown, SRD; the sample effect size, an estimate of the true effect size obtained from the sample; and a "critical value" of SRD (SRD*). The *critical effect size* defines the range of population effect sizes that would be considered of clinical or practical importance—that is, above the threshold of clinical significance, SRD > SRD* for a one-tailed test, |SRD| > SRD* for a two-tailed test. The value of SRD* is determined by what is learned in the background research.

For example, suppose we were evaluating two curricula A and B to be delivered in the first two school years, meant to result in an increase in IQ test scores in the fifth grade. If I were to tell you that with B (the "as usual" curriculum), the average IQ in third grade were 100 (±15), and that A is likely to raise that to 115 (±15), that would mean that δ = 1.0 ((115 − 100)/15)—in other words, SRD = .521 or NNT = 1.9 (a very large effect size). Few would argue that such an effect would not make a major difference, and that A should be seriously considered to replace B. But if I were

to tell you that actually A is likely to raise IQ to 100.15 (±15), that would mean that $\delta = 0.10$ (i.e., SRD = .056, NNT = 17.7). Now, instead of benefiting more than 1 of every 2 students (NNT = 1.9) by switching from B to A, we would benefit only 1 of every 18 students (NNT = 17.7). If the intervention carried any cost or inconvenience, the latter situation would hardly motivate such a switch—the two interventions would be considered, for practical purposes, equivalent. Somewhere between SRD = .056 (NNT = 17.7) and SRD = .521 (NNT = 1.9) is a point where potential users would begin to be motivated to switch to A. This could be determined by surveying those who will make the decision or examining the effect sizes of other interventions in the past in similar contexts that were ignored or adopted. The background research for the hypothesis to be tested should suggest, but does not prove, that the true effect size is of potential practical/clinical significance (i.e., greater than the critical effect size).

An Adequately Powered α-Level Test

An α-level significance test is *adequately powered* if the probability of a significant result is greater than a prespecified value (typically 80 percent) *for any effect size greater than the critical value*. In Figure 12.1, where the same critical value, SRD* = .3 (corresponding to $\delta = .5$), is applied in each case, the power of the two-tailed 5 percent test is 70 percent and that of the one-tailed 5 percent test is 80 percent. This simply illustrates that one typically needs a larger sample size for adequate power when testing a two- rather than a one-tailed hypothesis at the same significance level. However, the decision to use a one- or two-tailed test is not guided by power considerations, but by the *a priori* hypothesis.

Of the one-tailed 5 percent tests in Figure 12.2, the power of the *t* test to detect SRD* = .3 again is 80 percent, and those of the other tests based on dichotomization are 63 percent, 49 percent, or 35 percent for the same effect size, depending on where the cut-point is set for dichotomization. This illustrates a general principle: We almost always lose power when dichotomizing a reliable, valid scaled measure (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002), part of the reasoning behind and preference for the use of continuous variables. How much power we lose depends on where the cut-point is set. Here, maximal power occurs when the cut-point is exactly halfway between the two means. The further one moves away from that cut-point, the greater the loss in power. It is not unusual to have to double or triple the sample size to get adequate power with an ill-chosen dichotomization beyond what would have been adequate had the original scaled measure been used. Knowledge of such a principle is useful in choosing the test likely to be most powerful from among available valid tests.

If we chose to use the Mann-Whitney-Wilcoxon test on the original data, there would be a slight loss of power, but so slight that one would barely distinguish the power between the *t* test and the Mann-Whitney-Wilcoxon. Thus, another principle: If there is question about the validity of the *t* test, switching to a nonparametric test often (not always) means protecting the validity at only a slight reduction in power.

Once we limit consideration to valid 5-percent-level tests, and focus on the test within that group likely to have the greatest power (minimizing the necessary sample size, and thus the time, burden, and cost of the study), then and only then do we use the available tables, charts, computer programs, and so forth to compute the sample sizes (N, p) that will result in a valid, adequately powered test.

On Computations

While the present focus is not on computation of power, the subject cannot be completely avoided. First and foremost, it should be noted that all power computations are approximations! This is true because (1) each such computation is based on certain assumptions about the population (e.g., normal distributions, equal variances for the *t* test) that do not always hold exactly for the population actually sampled and (2) many approaches to computations are based on approximations to the non-null distributions.

Approaches to computations include mathematical derivations and formulas (not usually very helpful to researchers), tables and graphs in texts, and computer programs, including specialized statistical packages (e.g., Power and Precision), parts of general statistical packages (e.g., SAS or SPSS) and many free-ware programs to be found on the World Wide Web (e.g., Dattalo (2008, pp. 144–148)). All these approaches require that a decision be made on the particular test to be used and the significance level. They then use something indicating the critical effect size and something indicating sample size, to compute power. Exactly what form these "somethings" take and how the information is organized differ from one approach to another. What is more, the answers that result are not exactly the same from one approach to another.

For example, for a two-tailed t-test, Cohen (Cohen, 1988) first presents six tables, with one- and two-tailed tests at the 1 percent, 5 percent, and 10 percent levels (Tables 3.3.1 to 3.3.6), with rows defined by sample size per group and columns by δ, the standardized mean difference between the groups. The entries in the tables are levels of power. Then he provides another table (Table 3.4.1) in which rows are defined by desired power, columns again by δ, the entries of which are sample size per group. One or the other table may be more convenient for use in designing studies.

It should be noted that Cohen dealt explicitly only with the case of equal group sizes. However, if the proportion assigned to A is p, the total sample size needed is $n/(2p(1-p))$, where n is the sample size per group listed in Cohen's tables. Thus, when $p = 1/2$, the total sample size needed is 2n, n in each group. If $p = 1/4$, the total sample size per group needed is about $8n/3$, $2n/3$ in one group, 2n in the other. This fact is mentioned for two reasons: Every set of tables and charts, and often computer programs, have limitations. Once a user is familiar and comfortable with one approach, frequently such limitations can be worked out, as illustrated here.

In Cohen's book, each test (testing correlation coefficient, proportions, regression, etc.) is presented in a different chapter, with effect sizes differently defined and tables sometimes organized differently. In contrast, Kraemer and Thiemann (1987) present a grand total of only four tables for one-tailed and two-tailed tests, and 1 percent and 5 percent significance levels, with rows defined by an index called Δ, related to critical effect size, the columns defined by power, and the entries a parameter ν related to sample size. Then, for each different test, instruction is provided as to how to compute Δ and ν. This was done specifically so that users needed to learn to use only one table that would cover most common testing procedures.

For a 5 percent two-tailed *t* test, with two equal-sized groups ($p = 1/2$), for 80 percent power to detect a moderate critical effect size, $d^* = .5$, Cohen's table recommends 64 subjects per group, thus a total sample size of 128. In the same situation Kraemer and Thiemann recommend a total sample size of 133, 67 subjects per group.

Why the discrepancy? Which is more accurate? If the assumptions of the two-sample *t* test hold exactly, Cohen's answer is more accurate, because his calculations used the exact non-null distribution of the *t*-test statistic. Kraemer and Thiemann, on the other hand, used an approximation based on the exact distribution of the intraclass correlation coefficient, a distribution that can be used to approximate the distributions of many standard noncentral distributions, sacrificing some accuracy for ease of use. Similarly, Odeh and Fox (1991) based their charts on exact F-distribution, another approximation to the distributions of many standard noncentral distributions. This type of discrepancy is very common, not very large, but possibly confusing.

Other approaches include presenting computer programs to do specific power computations, or instructions on how to use standard statistical packages to do them (Aberson, 2010). Commercial software, as is also true of publications on power, undergoes considerable review and checking. However, free-ware often does not: some are quite accurate and convenient to use, and others are flawed. If free-ware is to be used, there should be some initial effort to check results against standards before committing to its use.

Otherwise, users may find one approach easier to use than another. For researchers, ease of use is likely to be the major factor in choosing one approach, since all the approaches (save a few free-ware approaches) are likely to be about equally accurate. As a practical matter, it is best to learn one approach thoroughly and to stick with that approach.

Execution as Planned

Once the study is designed and is documented to be a valid α-level test and to be adequately powered, and after approval and funding and whatever else is necessary to beginning the proposed study, the study is executed *as designed*, the test and effect size estimate performed *as planned*. The *p* value is computed, a statistic computed from the data equal to the probability of obtaining a sample effect size greater than the one actually observed with the study design, *when the null hypothesis is true*. Note that the significance level is a standard set *a priori*, whereas the *p* value is a statistic based on the data. If the *p* value is no greater than α, the result is described as *statistically significant at the* α *level*. Such a result results in rejection of the null hypothesis and thus support for the researchers' *a priori* hypothesis.

Reporting Results: Statistical Versus Practical Significance

The results in Figures 12.1 and 12.2 indicate that it is quite possible to have a statistically significant result that is of little or no clinical or practical significance, since, in an adequately powered study, for an effect below the critical effect size, the probability

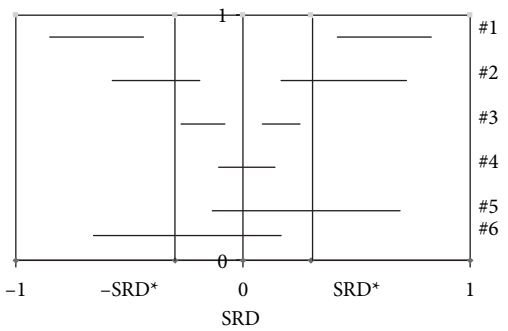


Figure 12.3 All possible configurations of 95 percent two-tailed confidence intervals for the effect size (here SRD) comparing two treatments A and B. Statistically significant results: #1, 2, 3. Clinical superiority: #1. Clinical equivalence: #3, 4. Failed study: #5, 6.

of a statistically significant result can still be quite large (e.g., here anywhere between 5 percent and 80 percent). Thus, reporting the *p* value should not be the end of the discussion, but the beginning. To continue, then, the sample effect size should also be reported, along with its confidence interval to specify the accuracy of the estimation. In Figure 12.3 are shown all the possible patterns of results for a two-tailed test:

- If the confidence interval does not contain any values consistent with the null hypothesis, the result is "statistically significant" at the set α level (repeating what would already be known from computation of the *p* value) (1, 2, 3 in Fig. 12.3).
- If the confidence interval contains no value greater than the critical value, *whether or not there is statistical significance,* such a result is said to prove "clinical or practical equivalence" (3, 4 in Fig. 12.3).
- Finally, if the confidence interval contains both values consistent with the null hypothesis (thus non-statistically significant) *and* values greater than the critical value (thus non-equivalent), it is a *failed study* (5, 6 in Fig. 12.3). With equipoise, before the study began, there was sincere doubt as to whether the researchers' theory was true or not. In a failed study, after the study is done, the sincere doubt remains; nothing has changed.

Summary

A valid, well-powered study is very unlikely to generate a false-positive result (less than 5 percent chance) and is unlikely to generate a false-negative result if the effect is of clinical/practical significance (less than 20 percent). A study does not fail simply because it does not generate a "statistically significant" result—after all, the null hypothesis may be true! It fails when the equipoise that existed before the study continues unshaken after the study.

Many incorrectly interpret a "non-statistically significant" result as "proving" the null hypothesis. However, absence of proof is not proof of absence. Most often, a nonsignificant result indicates inadequate design or poor measurement.

Many incorrectly interpret a "statistically significant" result as "proving" the researchers' theory. Such a result provides support for that theory, but the effect size may indicate that the hypothesis may be true but of little importance. In any case, independent replication or validation is usually required to establish a theory as scientific truth. The proliferation of poorly conceptualized and designed studies, studies that report invalid test results or that are underpowered, confuses and delays scientific progress.

Where Things Go Awry, and What to Do About It

Post Hoc Hypothesis Testing

When a hypothesis is generated by examination of the results in a dataset and is then tested on the same dataset, that is referred to as post hoc testing (see "The *A Priori* Hypothesis" above). The probabilities that define significance level and power pertain at the time the *a priori* hypothesis is proposed, before the data on which testing is to be done are available. The standard calculations of test statistics and *p* values are meant for that situation. When the data that stimulate the hypothesis overlap the data used to test that hypothesis, the *p* values are miscomputed. The probability of a false-positive result is then typically much greater than the preset significance level.

To take a simple illustration: For an *a priori* hypothesis that a certain coin is biased toward heads, a valid 5 percent test (not a very powerful one) would require that in five tosses of the coin, all five come up heads. When the null hypothesis of a fair coin is true, the probability of this result is $(1/2)^5 = .031 < .05$, less if the coin is biased toward tails, and thus is a valid 5 percent one-tailed test. However, suppose that, in the course of tossing coins, we happened to notice that one coin had produced four heads in a row, suggesting that the coin was biased toward heads, and only then did we propose the hypothesis that the coin was biased toward heads. The probability that we will now see five heads in a row with that coin, given that we

have already seen 4, is .50 for an unbiased coin, not less than .05! To validly test that hypothesis, we would have to initiate a *new* series of five tosses with that coin and see five heads in a row.

Many research papers have invalid *p* values arising from post hoc hypothesis testing sprinkled throughout the text like so much confetti. If readers simply ignored such tests, no harm is done. But it is often impossible to distinguish correct statistical testing from post hoc testing, and if the results of post hoc testing are taken seriously, problems may result.

Consider this example. The researchers report a clinical trial in which subjects are *randomly* assigned to interventions A and B with a well-justified hypothesis that A > B, with a well-designed study, and an adequately powered valid test of that hypothesis.

In the first table reporting results they present baseline descriptive statistics for the total sample. This information is crucial to defining the population to whom the conclusions of the trial apply. But then also reported are the baseline descriptive statistics for the A and B samples separately. This is unnecessary and redundant for descriptive purposes, since both are smaller *random* samples from the same population as is the total sample.

Randomization results in two random samples from the same population, not two matched samples. However, researchers often justify presenting the separate A and B descriptive statistics as a check on how well matched the two samples are. Were, for example, more of the boys assigned to the A group and more of the girls to the B group? Indeed, they then often go on to perform tests of the null hypothesis that the two samples are random samples from the same population for each of the baseline characteristics, even though clearly they know with certainty that this null hypothesis is true (see "Equipoise" above). After all, they themselves did the randomization! What is forgotten is that the computation of the *p* value comparing the two groups already includes consideration of all random samples, in some of which the number of girls and boys in the A and B groups is unequal.

Because the null hypothesis is here true, each such test has a 5 percent chance of being "statistically significant" (see "Significance Level" above). If *I* independent baseline characteristics were so tested, the probability that one or more will be statistically significant is $1 - (.95)^I$. Thus, if there are 15 or more such baseline measures, the probability of finding at least one statistically significant difference between the A and B random samples is greater than .54. But still, no harm is done—the reader can simply ignore these results.

However, what often happens is that researchers react to finding one or more statistically significant differences by replacing their *a priori* hypothesis (A > B in the total population) with a post hoc hypothesis that A > B within all subpopulations defined by whatever variables were seen to significantly differentiate A and B in this sample (i.e., "controlling" for those variables in testing). To do this, they often impose the unjustified assumption that the effect size in each subpopulation is exactly the same (using analysis of covariance, or omitting interactions from a multiple regression analysis). However, because the subsequent calculations are conditional on what was already seen in the sample, as well as on any assumptions made, the *p* values and the effect size estimates are now miscalculated. If the test had been run as originally planned, there would have been no such bias.

But, researchers protest, are we to ignore potentially important patterns we see in the data, simply because they are unexpected? Isaac Asimov is quoted as saying: "The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny' …." Of course, the answer is "No." After the statistical hypothesis testing is done *as originally proposed*, then the data should be explored to see whether any of the baseline variables (statistically significant or not) moderated the effect of treatment (Kraemer, Frank, & Kupfer, 2006; Kraemer, Kiernan, Essex, & Kupfer, 2008). If such moderators are detected, the existing literature would then be perused to augment support for a hypothesis of a moderated effect, providing rationale and justification for a *subsequent* trial to test a more complex hypothesis (see "Theoretical Rationale and Empirical Justification" above), appropriately designed and powered to test that hypothesis (see "A Valid Statistical Test," "Significance Level," "Effect Size," "Power," and "The Critical Effect Size" above).

Cherry Picking

One quite common and egregious error is often referred to as "cherry picking." Here the *a priori* hypothesis posits an effect on an outcome measure (say, A > B for IQ test scores in the fifth grade) (see "The *A Priori* Hypothesis" above). However, when the study is executed and the testing done, no statistically significant difference in that specified outcome measure is found. Then researchers examine

subtest scores on that IQ test, or other outcomes, such as behavioral problems, school attendance, and so forth. Even if the null hypothesis (here A > B) is true for all outcomes, testing multiple outcomes will inevitably lead to one or more false-positive results (see "Post Hoc Hypothesis Testing" above). "Cherry picking" refers to the situation in which the researchers replace their original outcome measure with one found statistically significant. It is difficult to spot "cherry picking" in published studies. However, a study (Chan, Hrobjartsson, Haahr, Gotzsche, & Altman, 2004) that compared protocols for studies with their published reports found that "62 percent of trials had at least 1 primary outcome that was changed, introduced, or omitted" (p. 2457). This situation is a major reason for the requirement for registration of randomized clinical trials (DeAngelis et al., 2005) and one major source of false-positive results in the research literature. In terms of externally funded research, your proposal must specify the primary dependent variable, and the decision cannot be changed during or after the study.

Multiple Testing

From the above discussion, it might sound as if each study can test one and only one hypothesis. That clearly is not true, and it would be wasteful of the time, effort, and resources each hypothesis-testing study requires. It is certainly true that if there were only one hypothesis to be tested, every design, measurement, and analytic decision could be focused on what is optimal for that one research question. As soon as there is more than one hypothesis to be tested, we must often compromise in designing the optimal study for one hypothesis to ensure the validity of testing another. The more hypotheses, the greater the potential number of compromises and the poorer the quality of the study for any individual hypothesis. Thus, one hypothesis per study is clearly not ideal, but too many hypotheses per study can be disastrous.

To have multiple tests of essentially *one* hypothesis is clearly problematic. For example, if one primary outcome is the IQ test score in the fifth grade, IQ test scores in the third and fourth grades, since they relate to the same underlying construct, should not be tested as separate hypotheses. One might instead hypothesize that A > B in promoting a better trajectory of IQ test scores between grades three and five, and combine these into one test (more powerful than would be the test on the fifth-grade score alone). In general, if multiple hypotheses are to be considered, they should be conceptually independent; in other words, knowing the correct answer on any one provides no clue as to the answer on any other.

When the multiple hypotheses to be tested are conceptually independent, an argument can be made that each hypothesis can be treated as if tested in a different study. Thus, for example, in the A-versus-B comparison, one might (1) test a hypothesis in the cross-sectional prerandomization data concerning the association between IQ test scores in the first grade with teacher-reported behavioral problems; (2) test a hypothesis in the B group only (usual curriculum) as to how well first-grade IQ test scores predict fifth-grade IQ scores; and (3) test the hypothesis already discussed comparing A versus B on fifth-grade test scores. The challenge is to include no hypothesis that would require compromising the testing of another hypothesis, and to ensure adequate power for testing the weakest of the hypotheses.

An alternative approach that researchers often prefer is to divide the conceptually distinct *a priori* hypotheses to be tested into those that are primary (often only one) and those that are secondary. All are well justified and *a priori*, and a valid statistical test is proposed for each. However, the study is designed to have adequate power to test only the primary hypothesis. It would then be understood that there may be a major risk of failing to get any answer for some of the secondary hypotheses, for the success of the study focuses on avoiding failure for the primary hypothesis.

The Nonspecific Hypothesis

Occasionally researchers claim only one hypothesis, such as "Some gene relates to some mental disorder," but propose a study in which 1,000 gene loci are each correlated with 10 mental disorders. While the hypotheses can here be summarized in one brief statement, there are actually 10,000 specific hypotheses, for each of which rationale and justification should be provided. A specific hypothesis would be "The 5-HTT gene is related to the onset of major depressive disorder," supported by information about the 5-HTT gene and about major depressive disorder and giving indications as to why this particular association might exist and be important. With 10,000 associations to be tested, such justification for each is clearly impossible.

The 10,000 hypotheses here, in any case, are not conceptually distinct. Genes are often linked; mental disorders are often comorbid; even independent genes may interact with each other to influence

mental disorders or the comorbidities among them. Thus, 10,000 or more statistical tests are run, with no individual rationale and justification, with no particular concern as to whether the study is well powered for each hypothesis or not.

It is well recognized that if the 5 percent significance level is used for each such test, the probability of one or more "statistically significant" results would be much larger than 5 percent, a proliferation of false-positive results (see "Post Hoc Hypothesis Testing" above). Thus, it is usually recommended that we protect against this by adjusting significance level in one way or another. For example, the significance level for each individual test might be set at .05/T, where T is the number of tests, .05/10,000 = .000005. Then it can be shown that, if there were no associations between any of the genes and mental disorders, the probability of one or more statistically significant results is less than .05 (Bonferroni correction; Bailey, 1977; Benjamini & Hochberg, 1995). However, the null hypothesis now rejected posits complete absence of associations and thus supports the conclusion that *somewhere* there is a nonrandom association. It is not true that the specific gene–disorder pairs that have individual *p* values less than .05/M are the true positives, or that the others are true negatives. Moreover, adequate power at the .05/M significance level requires many, many more subjects, and thus greatly increases the cost of the study. In most cases, adequate power and effect sizes are not even considered. Thus, many of the "statistically significant" pairs may have effect sizes indicating trivial association, while many of the "non-statistically significant" pairs may indicate strong association but inadequate power to detect that association. Finally, the results of one such study are seldom replicated in a second.

In a cutting-edge research area, as is the gene–mental disorder link, it is reasonable to propose an exploratory (hypothesis-generating) study in which indeed 1,000 genes, and perhaps even various combinations of those genes, may be correlated with 10 mental disorders, but *statistical hypothesis testing should not there be done*. Interpretable effect sizes indicating the strength of association between each gene and mental disorder and their confidence intervals might be computed, and these used to identify "hot spots"—certain gene–mental disorder pairs that seem particularly strongly associated. Then any other previous research relating to a particular "hot spot" would be compiled, and if now there seems rationale and justification for a specific hypothesis relating that genotype (perhaps involving multiple gene loci) and that mental disorder, this association could then be formally tested in a subsequent study designed for that purpose. The problem lies not in doing a study correlating 1,000 genes and 10 mental disorders, but in misusing statistical hypothesis testing to try to draw conclusions from such a study, rather than using the data from such a study to generate strong hypotheses to be tested in future valid statistical hypothesis-testing studies well designed for the purpose.

Confusion Between the True Effect Size, the Critical Effect Size, and the Estimated Effect Size

The *true effect size* related to any hypothesis is unknown, and is never known exactly (see "Effect Size" above). It is estimated by the *sample effect size* based on the study done to test a hypothesis concerning that effect size, with a certain margin of error indicated by its confidence interval (see "Effect Size" above). The *critical effect size* is a standard against which the true or sample effect size is assessed for clinical/practical significance (see "The Critical Effect Size" above). A study is an adequately powered 5 percent-level-test probability of rejecting the null hypothesis *when the true effect size consistent with the null hypothesis* is less than 5 percent, and if the probability of rejecting the null hypothesis *when the true effect size exceeds the critical effect size* is greater than 80 percent (see the sections "Significance Level" and "The Critical Effect Size" above). Whether the true effect size exceeds the critical effect size or not is then indicated by comparisons between the sample effect size and the critical effect size (Fig. 12.4). Confusion of the three is a source of many problems.

For example, researchers often try to guess the true effect size and equate that to the critical effect

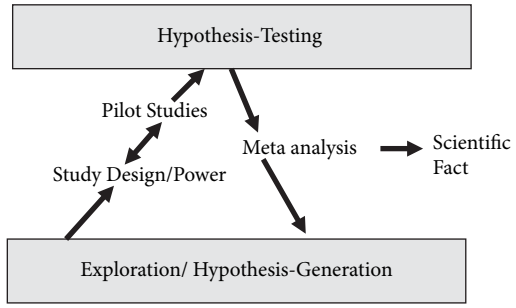


Figure 12.4 The ideal structure underlying hypothesis-testing studies. Power is an issue only in the study design phase for hypothesis-testing studies, not for exploratory studies, pilot studies, or meta-analyses.

size, as if they believed that the purpose of a research study is to get a statistically significant result rather than to find out whether the hypothesis is true and important or not. It is crucial to remember that the researchers' hypothesis may be false or it may be true but trivial (see "Execution as Planned" above), in which case a non-statistically significant result is the preferable result.

Similarly, researchers often do a small "pilot study" specifically to get an estimate of the true effect size, which is then used as the critical value for power calculations (Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006). With a small study, the error of estimation of the true effect size is large, and gross under- and over-estimation is very likely. If the true effect size is underestimated, the proposed study may be aborted even if the true effect size is large and important, either because the sample size necessary is too large to be feasible, or because the result is discouraging. If the true effect size is overestimated, the proposed study may be funded and executed, but it is likely to be a failed study because it will be underpowered. In short, using an estimated effect size from a small study as the critical effect size in designing a proposed study is a serious error, and likely the source of many underpowered, failed studies.

Realistically, trying to determine the critical value in any context is very difficult. For randomized clinical trials, Cohen (1988) suggested that $\delta = .2, .5$, and .8 were "small," "medium," and "large" effects (corresponding to SRD = .1, .3, .4 or to NNT = 9, 4, and 2). However, the NNT for the Salk vaccine to prevent polio is NNT = 2,500, considered a very strong effect (because most children do not get polio whether or not they are vaccinated). It has been suggested that when comparing an active drug against an inactive placebo for treatment (not prevention) of a disorder, the critical effect size is likely to be around .8, while when comparing two active drugs against each other, it might be around .2., and the same holds for comparisons of psychological treatments.

Similarly, Cohen suggested that correlation coefficients $\rho = .1, .3$, and .5 be considered "small," "medium," and "large." However, a test–retest reliability coefficient of .3 would be considered low and unacceptable for many test scores, while a correlation between a gene and a mental disorder of .1 might be considered quite high (correlations between binary measures, particularly when one or the other describes a rare event, are often very small).

Cohen warned, in suggesting these standards, that they should not be reified but should serve as starting points for the discussion of what the critical effect size is in the specific context of the research. However, such discussions still seem to be rare in designing hypothesis-testing studies, and are sorely needed.

Inadequate and Incomplete Reporting of Results

When the results are reported only with *p* values, or even worse with asterisks representing $p < .05$, $p < .01$, there may be support for the hypothesis but no indication of the scientific importance of the result (clinical/practical significance). It is not yet routine to report interpretable effect sizes and their confidence intervals, and in the absence of these, issues related to clinical/practical significance remain unaddressed, much less resolved. This situation is particularly problematic in epidemiological studies, where the sample sizes are often in the thousands. Then risk factors are identified as statistically significant that are so weak that, even were the risk factor a cause of the disorder, and even if the risk factor could be removed completely by intervention, the incidence/prevalence of the disorder would barely be affected. In the meantime, the intervention delivered to those who do not need it, might be, at the very least, a waste of time, effort, and money, and at the very most, also harmful, increasing the incidence/prevalence of other disorders.

Backward Power Calculations

For a 5-percent-level *t* test, if we knew N, p, and the critical effect size, we could use existing standard methods to compute power at the critical effect size. This calculation is done to determine whether or not what is proposed is adequately powered. If one knew p, and the desired power, and the critical effect size, we could compute N for an adequately powered study. Both of these calculations are valuable in designing a study, for one can negotiate design decisions, choice of analytic strategy, N, and p to generate adequate power.

It is also technically true that one could compute an effect size for which power would be adequate (say 80 percent), knowing N and p, but the effect size so calculated is *not* the critical effect size. Because the tendency to do this "backward" power calculation is stronger when the proposed sample size is small, the effect size so calculated is usually much stronger than the critical effect size. To show that with 20 subjects per group, one would have 80

percent power to detect an effect size far beyond what is truly a critical effect size is merely a waste of time and resources, and most commonly results in failed studies.

Post Hoc Power

Backward power calculations are done *a priori*, but there is also a problem with so-called post hoc power computations. After a hypothesis-testing study is done, if there is no question of validity and the result is a statistically significant effect, Lincoln Moses' Principle of the Dull Axe applies: "If the tree fell down, the axe was sharp enough!" (Goldstein, 1964, p. 62). If the result is not statistically significant, then the design was faulty or the measurement poor ("the axe not sharp enough"), or the rationale and justification were misleading (type III error, the failure to ask the right questions in the first place), or the researchers simply had bad luck (80 percent power still means up to a 20 percent chance of a nonsignificant result!).

In such failed studies, post hoc power calculations are sometimes done using the sample effect size as if it were the critical effect size. Researchers seem to compute post hoc power to convey the idea that they have not committed a type III error, that it was "only" a type II error (inadequate power), and thus their hypothesis remains viable. However, the correct power calculation is still that done *a priori* using the critical effect size. That does not change because the study has failed. Moreover, doing such a calculation after the study is done, using the sample effect size as the critical value, doesn't change the uncertainty as to the cause of a failed study and doesn't affect equipoise (Levine & Ensome, 2001).

Instead, when the study is done, as long as there is no question of its validity, the sample effect size might be combined in a meta-analysis (Cooper & Hedges, 1994; Hedges & Olkin, 1985) with other valid estimates of the same population effect size to see whether consensus has been reached either that the hypothesis is true and of clinical significance using the critical effect size as the criterion (Situation 1 in Fig. 12.3) or that the hypothesis is no longer likely to be clinically significant (Situations 3 and 4 in Fig. 12.3). It may be that no single study finds a statistically significant result, but the consensus of multiple such studies does, and may disclose a definitive result (that A is clinically superior to or equivalent to B). Otherwise, a further study, better designed, is warranted to test the hypothesis, for the rationale and justification continue to apply, as does equipoise.

The Structure and Context of Statistical Hypothesis Testing

From this discussion it may be clear that there is a logical sequence of efforts leading up to and out from valid and adequately powered hypothesis-testing studies, a sequence that leads to fewer failed studies, fewer false positives in the research literature, conservation of resources necessary to execute studies, and consequently faster research progress (Fig. 12.4).

To make this explicit, let us start with exploratory, hypothesis-generating studies that provide rationale and justification for hypothesis-testing studies. Hypothesis testing should not be done in such studies, and no *p* values should be reported, for there are no *a priori* hypotheses. The purpose here is to generate strong hypotheses and the information necessary to design adequately powered, valid such studies.

Many of the problems discussed above arise because valuable, well-done, hypothesis-generating studies are misrepresented as hypothesis-testing studies. This misrepresentation, in turn, results in part from a bias among proposal reviewers and reviewers and editors of papers submitted for publication against such hypothesis-generating studies. Such studies are often referred to by derogatory terms such as "fishing expeditions," "data dredging," "torturing the data until they confess," and so forth. The result is that researchers seeking to get valuable information into the research literature misrepresent their studies as hypothesis testing, thus generating "conclusions" that are very likely to be false positives (type I errors).

In the absence of well-done exploratory studies, the hypotheses tested in hypothesis-testing studies are often weak (type III errors), and the designs proposed often are flawed (type I and II errors) because the information needed for strong hypotheses and valid, powerful designs is lacking.

Clearly two changes are needed. First must be a greater receptivity to *well-done* exploratory studies, both in proposal and paper review. Then data sharing of datasets from well-done hypothesis-testing studies, after the researchers have completed their *a priori* proposed use of the data, is essential to facilitate exploratory work by all researchers. In such studies, *p* values should not be reported; the focus should be on generating hypotheses for future testing.

Next comes the process of designing a hypothesis-testing study, based on the rationale and justification in the exploratory studies, and supported by information from those studies necessary to good design, measurement, and analytic decisions. This design phase for a hypothesis-testing study is the one and only phase at which power considerations pertain, and here they are paramount. As Light and colleagues say: "You can't fix by analysis what you muddle by design!"(Light, Singer, & Willett, 1990, p. v).

However, there are often tactics proposed in the design of such a study that are not necessarily feasible or optimal in the milieu in which the study is to be done. Is it, for example, actually possible to recruit 100 eligible subjects per year as proposed? Will patients accept randomization to the treatments proposed? If there are five different IQ tests available, which one should be used?

To avoid the unfortunate situation in which, after many subjects have already entered and been studied, the researchers find that the proposed procedures are impossible to follow consistently or are contraindicated, a pilot study can be conducted. A *pilot study* is a small study, done as a precursor to a full-scale hypothesis-testing study, to evaluate the feasibility of any aspects of the proposal for which there is question. Not only are pilot studies small, they are often different from the study eventually proposed, for errors detected in the pilot study motivate changes in the protocol of the proposed study. Thus pilot studies do not test hypotheses or estimate effect sizes. No *p* values should be reported. In particular, pilot studies should not be used to estimate the true effect size to be used as the critical effect size in designing the proposed study (Kraemer, Mintz, et al., 2006). However, pilot studies are often essential to a *successful* subsequent hypothesis-testing study.

The term "pilot study" has unfortunately come to be used to describe not a feasibility study preparing for a hypothesis-testing study but rather any small, badly designed, or underpowered study. At the same time, many legitimate pilot studies are evaluated as if they were hypothesis-testing studies. Proposers of pilot studies have had the experience of being asked by reviewers whether the proposers had done a pilot study leading to the present pilot study! Reviewers often inappropriately ask for hypotheses, tests, and power calculations. With multiple testing and "cherry picking" in such studies, researchers often report invalid statistically significant results as conclusions. At the same time, the feasibility problems that the legitimate pilot studies are supposed to avert continue to occur, often leading to very costly failed hypothesis-testing studies.

Once the hypothesis-testing study is designed, and, if necessary, "debugged" using the results of a pilot study, the study should then be executed *as proposed* to address the hypotheses. When a hypothesis-testing study is concluded, meta-analysis might be used to combine the new estimate of the effect size with other valid estimates of the same population parameter from any earlier studies to see if the result is now definitive (a scientific fact) (Lau et al., 1992), in which case equipoise no longer pertains. It is generally accepted that independent replication or validation is necessary to establish a scientific fact. Thus one study, no matter how valid and adequately powered, how well executed and analyzed, cannot establish a scientific fact, but it will seldom take more than three to five valid, adequately powered studies to do so. However, it will take far more valid but inadequately powered studies (Kraemer, Gardner, Brooks, & Yesavage, 1998), and including invalid studies will clearly confuse the results and make it very difficult to ever reach correct consensus.

Subsequently, the data from the hypothesis-testing study should be made available for exploration, not only to check and to illuminate the results on the hypotheses already tested, but also to generate hypotheses either broader or deeper than the ones in this study, for testing in future studies.

Thus, for well-conceived, well-executed, successful hypothesis testing, the process comes full cycle (see Fig. 12.2). We start with hypothesis-generating studies and we finish with hypothesis-generating studies, with valid, adequately powered hypothesis-testing studies at the peak of the research process.

Power computations matter in the context of designing hypothesis testing of well-justified *a priori* hypotheses but have no application otherwise, not in exploratory studies, not in pilot studies, not in meta-analysis. Imposing hypothesis testing in exploratory or pilot studies, using it in the absence of well-justified and *a priori* hypotheses, using post hoc testing, "cherry picking" results, and so forth are all serious misuses of a valuable and unique tool, statistical hypothesis testing, and have resulted in a proliferation of false-positive and false-negative results, confusing the scientific literature and slowing scientific progress. To avoid such problems in future research, we need not more but far less statistical hypothesis testing, with such testing focused on valid, adequately powered tests, with *p* values and sample effect sizes and their confidence intervals reported.

References

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Routledge Taylor & Francis Group.

Bailey, B. J. R. (1977). Tables of the Bonferroni t statistic. *Journal of the American Statistical Association*, *72*, 469–478.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*(1), 289–300.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcome in randomized trials. *Journal of the American Medical Association*, *291*(20), 2457–2465.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Dattalo, P. (2008). *Determining sample size*. Oxford: Oxford University Press.

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2005). Is this clinical trial fully registered? *Journal of the American Medical Association , 293*(23), 2927–2929.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine*, *317*, 141–145.

Goldstein, A. (1964). *Biostatistics: an introductory text*. New York: The McMillan Company.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press Inc.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3–7.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 285–299.

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), 1–4.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). The advantages of excluding under-powered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, *3*, 23–31.

Kraemer, H. C., Kiernan, M., Essex, M. J., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2), S101–S108.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*(11), 990–996.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, *63*(5), 484–489.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage Publications.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*(448), 1372–1381.

Lau, J., Elliott, M. A., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine*, *327*, 248–254.

Levine, M., & Ensome, M. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, *21*(4), 405–409.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design*. Cambridge, MA: Harvard University Press.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40.

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. New York: Routledge Taylor & Francis Group.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Odeh, R. E., & Fox, M. (1991). *Sample size choice: Charts for experiments with linear models*. New York: Marcel Dekker, Inc.

Paulos, J. A. (1988). *Innumeracy*. New York: Hill and Wang.

Shearer-Underhill, C., & Marker, C. (2010). The use of number needed to treat (NNT) in randomized clinical trials in psychological treatment. *Clinical Psychology: Science and Practice*, *17*, 41–48.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*(1), 1–2.

CHAPTER

13 Multiple Regression: The Basics and Beyond for Clinical Scientists

Stephen G. West, Leona S. Aiken, Heining Cham, *and* Yu Liu

Abstract

This chapter presents a broad overview of multiple regression (MR), psychology's data analytic workhorse. MR is a statistical method for investigating the relationship between categorical, quantitative, or both types of independent variables and a single dependent variable. An initial literature review documents the broad use of MR in leading journals in clinical psychology. The chapter then provides an understanding of the fundamentals of simple linear regression models, followed by more complex models involving nonlinear effects and interactions. The chapter presents solutions to several potential problems that arise in MR, including missing data, data collected from groups (multilevel modeling), data collected from individuals over time (growth curve models, generalized estimating equations), and noncontinuous dependent variables including binary outcomes, unordered or ordered categories, and counts (generalized linear model). Throughout, the chapter offers advice for clinical researchers on problems that commonly arise in conducting and interpreting MR analyses.

Key Words: Multiple regression, interaction, curvilinear, logistic regression, generalized linear model, growth model, multilevel model

Multiple regression (MR) is a very flexible data analytic approach. Indeed, because of its flexibility and wide use, MR has been called "psychology's data analytic workhorse" and "everyday data analysis." In its most basic form MR addresses questions about the linear relationship between one or more independent variables (IVs) and a single dependent variable (DV). The IVs can include any combination of qualitative variables (e.g., gender, diagnostic category, experimental treatment conditions), quantitative variables (age, IQ, score on a scale such as the Beck Depression Inventory [BDI], EEG measure), or both qualitative and quantitative variables. The DV is a single continuous dependent variable (e.g., scores on an assessment, measure of marital happiness). Although the form of the relationship between the IVs and DV is most commonly represented as linear, MR can also address a variety of more complex curvilinear relationships and interactive relationships, with the latter involving the combined effects of two or more IVs. Traditional analysis of variance (ANOVA) and correlational analyses can be seen as special cases of MR.

In recent years more advanced analyses that build on the foundation of MR are increasingly being used by clinical scientists, as we will document. These analyses extend the basic logic of MR but address specific issues that arise when the research involves DVs that are not continuous or data structures in which the observations are not independent. The generalized linear model addresses noncontinuous DVs. Examples of specific applications include logistic regression, different variants of which are appropriate for binary dependent variables (e.g., diagnosis, case vs. non-case), ordinal outcomes (e.g., severity of problem: mild, moderate, severe), and Poisson

regression and negative binomial regression, which are appropriate for counts (e.g., number of aggressive episodes in a fixed period of time). Extensions of MR also address nonindependence that arises from studies in which participants are clustered into groups (e.g., group therapy, research on families) or repeated assessment of the same set of participants in longitudinal research. Multilevel modeling (a.k.a. hierarchical linear models) addresses clustering into groups. Multilevel modeling, growth trajectory models, and generalized estimating equations address repeated assessments of the DV over time. Following in the tradition of Cohen, Cohen, West, and Aiken (2003) and Fox (2008), we provide a broad perspective on MR, including material that is often discussed under the rubrics of the general linear model, the generalized linear model, and multilevel modeling.

Our goal is to provide a perspective on MR that will be useful to both beginning and advanced clinical researchers. We begin this chapter with a brief survey of two leading journals in abnormal and clinical psychology to provide a snapshot of recent use of MR by clinical researchers, information on which we will periodically draw throughout the chapter. We then review the basics of MR with a focus on issues of particular relevance to clinical researchers: basic linear models, diagnostic procedures that detect violations of assumptions, and curvilinear effects and interactions. Following the development of this foundation, we consider several advanced issues that arise in both basic research in abnormal psychology and intervention research in clinical psychology. Topics include missing data, multilevel models, growth curve models and models for repeated measures, and noncontinuous outcomes. Throughout the chapter we will identify issues that lead to common problems in MR analysis for clinical researchers.

Survey of Statistical Methods in the Recent Abnormal and Clinical Literature

We conducted a survey of the statistical methods reported in the *Journal of Abnormal Psychology* (*JAP*) and the *Journal of Consulting and Clinical Psychology* (*JCCP*) in the second half of 2010. As a brief overview, in *JAP* 59 separate studies were reported in the 56 empirical research articles in issues 2 through 4. More than one analysis could be reported in each study. Of the 59 studies, 43 (73 percent) used MR, 2 (3 percent) additional studies reported repeated-measures ANOVA, and 24 (41 percent) used an extension of MR. Extensions included multilevel modeling (6 studies, 10 percent), latent trajectory growth models (2, 3 percent), generalized estimating equations (6, 10 percent), and the generalized linear model (13 total, 22 percent), including logistic regression (12), Poisson (1) and negative binomial regression (2) for counts, multinomial logistic regression (1), and ordinal logistic regression (2). Sample sizes in these studies ranged from 28 to 8,580 (median = 177). There were 8 studies of longitudinal growth with 4 to 10 (median = 8) observations per participant, plus one experience sampling study with 6 random ratings per day across 21 days (Nisenbaum, Links, Eynan, & Heisel, 2010).

In *JCCP* 46 separate studies (excluding 4 meta-analyses) were reported in 49 articles in issues 4 through 6. Of the 46 studies, 35 (76 percent) used MR,[1] 6 (13 percent) additional studies reported repeated-measures ANOVA, and 29 used an extension of MR (63 percent). Extensions included multilevel modeling (8 studies, 17 percent), latent trajectory growth models[2] (14, 30 percent), generalized estimating equations (3, 7 percent), and the generalized linear model (14 total, 30 percent), including logistic regression (6), Poisson (2) and negative binomial regression (1), multinomial logistic regression (2), ordinal logistic regression (1), and survival analysis (3). Sample sizes ranged from 19 to 1,822 (median = 208). In all, 35 studies were longitudinal with 2 to 18 observations per each participant (median = 3).

Taken together, these findings indicate that MR continues to be the data analytic workhorse in studies of basic abnormal psychology, emphasized in *JAP*, and in clinical intervention studies, emphasized in *JCCP*. However, they also point to the increased use of more complex models that are extensions of basic MR. The more complex models require a solid knowledge of basic MR as a foundation for the proper interpretation of their results.

Multiple Regression: The Basics

Linear Model

Although MR may be used for description or prediction, it is typically used in abnormal and clinical psychology to test a specific hypothesis developed from theory or prior research. The hypothesis identifies (a) the IVs that are expected to relate to the DV and ideally (b) the form of the relationship between the IVs and the DV. The researcher may also wish to control statistically for other background variables (e.g., education, gender, baseline measure). MR provides the appropriate statistical machinery to address these issues. The model being tested requires careful consideration by the analyst. The IVs must be thoughtfully chosen and the nature

of the relationships among the IVs and of the IVs to the DVs, both alone and in combination with other IVs, must be considered. In this section we focus on the interpretation of basic MR results in the simple linear regression model. We also illustrate the importance of careful selection of IVs, as this can be a source of results that at first glance are surprising. Later we consider more complicated nonlinear relationships, both curvilinear and interactive.

Consider the following hypothetical example. A Ph.D. student in clinical psychology hypothesizes that the stress associated with the job and a prior diagnosis of depression will each predict later poor psychological functioning in the workplace. She identifies 100 (50 males, 50 females) newly employed individuals who have previously received a clinical diagnosis and a control group of 100 (50 males, 50 females) newly employed individuals who had never received a diagnosis. Each participant's *work supervisor* fills out an 11-point scale of job stress (*Stress*) reflecting the demands of the employee's position, where 0 = no demands and 10 = maximum possible demands. Each participant is measured on the DV, an 11-point scale of *Poor Workplace Functioning*, where 0 = excellent workplace functioning and 10 = extremely poor workplace functioning. Finally, the student collects measures of participant *Gender* (Male = 1; Female = 0) and *IQ* as background variables that should be statistically controlled in the analysis. The researcher hypothesizes that job stress and prior diagnosis of depression will each contribute to poor psychological functioning in the workplace, even after the participant variables of participant IQ and gender have been controlled. The student proceeds with a systematic set of analyses that investigate the effects of each variable alone and in combination with the others.

(A) *Effects of Diagnosis*. The student initially wishes to test the hypothesis that a prior diagnosis of depression predicts poor workplace functioning. Diagnosis is treated as a dummy variable, labeled *Diagnosis* (see later description of dummy coding). A score of 1 is assigned to *Diagnosis* if the employee has a previous diagnosis of depression; a score of 0 is assigned if the employee has never had a diagnosis. The group with a score of 0 serves as the reference or comparison group. To compare the two groups, she runs a regression with prior diagnosis as the predictor:

$$Y = b_0 + b_1 Diagnosis + e \quad (1a)$$

$$\text{or, } \hat{Y} = b_0 + b_1 Diagnosis \quad (1b)$$

Equations (1a) and (1b) are identical, except that Equation (1a) focuses on the observed Y (the actual score of each person on *Poor Workplace Functioning*) and includes the residual e. Equation (1b) focuses on predicted Y ($\hat{Y}$), predicted *Poor Workplace Functioning*, and does not include the residual. To explain further, Equation (1a) shows that *Diagnosis* is related to the observed score Y, but that the relationship is not perfect. A portion of each person's observed Y score is unrelated to depression *Diagnosis*; this portion is represented in the residual (e), the difference between the predicted value $\hat{Y}$ and the observed value Y for each participant, $Y - \hat{Y}$. Each participant will have a different residual. Equation (1b) represents only the portion of the person's Y score ($\hat{Y}$) that is predicted from Diagnosis. Table 13.1A shows results in this hypothetical example. Figure 13.1A provides a graphical depiction of the relationship between diagnosis and psychological functioning. In Equation (1a), Y (here, *Poor Workplace Functioning*) is the observed value on the DV, b_0 is the predicted value of Y in the group coded 0 (comparison group), and b_1 is the difference between the predicted value of the DV for the participants in the diagnosis minus that for the comparison group. *Diagnosis* is a group variable, so $b_0 = 4.07$ is the mean of the comparison group on the DV, $\hat{Y}_{Comparison}$, and $b_0 + b_1 = 4.07 + 0.86 = 4.93$ is the mean of the diagnosis group on the DV, $\hat{Y}_{Diagnosis}$. Finally, Table 13.1A reports $R^2 = 0.18$, the proportion of variation in Y accounted for by *Diagnosis*. R^2 may be equivalently interpreted as the proportion of the total variation in Y accounted for by the IVs or the squared correlation between the predicted values $\hat{Y}$ and the observed values Y—that is, $(r_{Y\hat{Y}})^2$.

(B) *Effects of Job Stress*. High levels of job stress are expected to be positively related to poorer psychological functioning in the workplace. The student estimates a second regression equation to test this hypothesis:

$$\hat{Y} = b_0 + b_1 Stress \quad (2)$$

Table 13.1B shows the results and Figure 13.1B provides a graphical depiction of the relationship between work *Stress* and *Poor Workplace Functioning*. In Equation (2), $\hat{Y}$ is again the predicted value on the DV, $b_0 = 3.25$ is the intercept, the predicted value of Y when Stress = 0. $b_1 = 0.25$ is the increase in Y (*Poor Workplace Functioning*) for each 1-unit increase in *Stress*. When the IV is a quantitative variable, it is often useful to report the standardized regression coefficient $b_1^* = 0.40$, which represents the number of standard

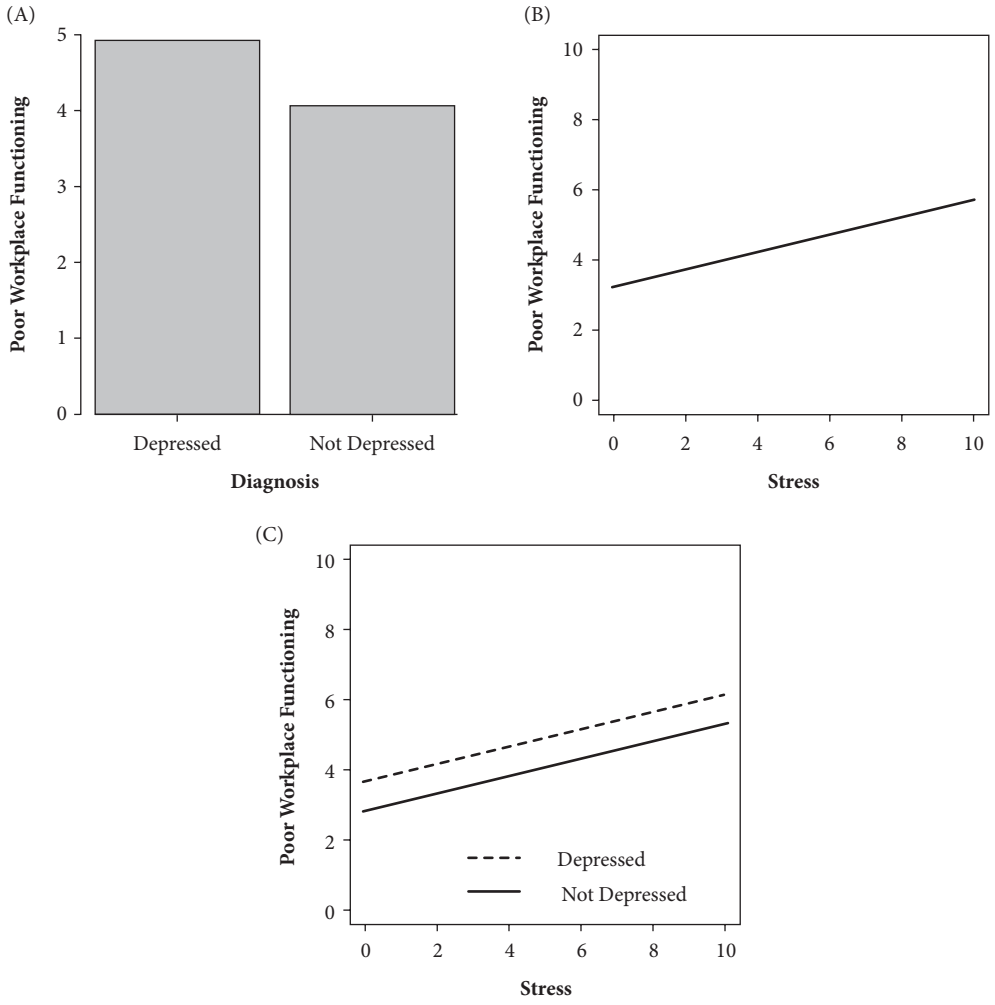


Figure 13.1 *Poor Workplace Functioning* predicted from (**A**) prior *Diagnosis* of depression, (**B**) job *Stress*, and (**C**) the additive effects of both independent variables.

deviations Y increases for a 1-SD change in the IV.[3] Again, we report $R^2 = 0.16$, here the proportion of the total variation in Y accounted for by *Stress*.

(C) *Unique Effects of Diagnosis and Job Stress*. The student now wishes to study the unique effect of *Diagnosis* and the unique effect of *Stress* with the other IV controlled. To do this, she estimates a regression equation that includes both IVs:

$$\hat{Y} = b_0 + b_1 Diagnosis + b_2 Stress \qquad (3)$$

Table 13.1C shows the results and Figure 13.1C provides a graphical depiction of the relationship between *Diagnosis*, *Stress*, and *Poor Workplace Functioning*. In Equation (3), $b_0 = 2.82$ is the intercept, the predicted value of Y when both *Diagnosis* = 0 (comparison group) and Stress = 0, $b_1 = 0.86$ is the difference in the mean levels of *Poor Workplace Functioning* in the diagnosis minus the comparison groups, holding stress constant at a fixed value, and $b_2 = 0.25$ is the increase in *Poor Workplace Functioning* holding *Diagnosis* constant at a fixed value. As can be seen in Figure 13.1C, the regression lines for the diagnosis and comparison groups are parallel in this model.

(D) *Effects of Diagnosis and Stress Over and Above Background Variables*. Very often researchers will be concerned that background covariates might be confounding variables that could account for the observed relationships between the IVs and the DV. About half of the articles reviewed included one or more baseline covariates (most often one or more of gender, age, ethnicity, and a baseline clinical measure). Although this practice is often helpful in

Table 13.1 Analyses of Regression Models

(A) Regression Model of Equation (1) [$R^2 = 0.18$]				
IV	**Unstandardized *b***	**Standard Error (*b*)**	**Standardized *b****	***t*(198)**
Intercept	4.07	0.091		44.91**
Diagnosis	0.86	0.128		6.70**
(B) Regression Model of Equation (2) [$R^2 = 0.16$]				
IV	**Unstandardized *b***	**Standard Error (*b*)**	**Standardized *b****	***t*(198)**
Intercept	3.25	0.214		15.21**
Stress	0.25	0.041	0.40	6.14**
(C) Regression Model of Equation (3) [$R^2 = 0.34$]				
IV	**Unstandardized *b***	**Standard Error (*b*)**	**Standardized *b****	***t*(197)**
Intercept	2.82	0.198		14.26**
Diagnosis	0.86	0.115		7.46**
Stress	0.25	0.036		6.94**
(D) Regression Model of Equation (4) [$R^2 = 0.41$]				
IV	**Unstandardized *b***	**Standard Error (*b*)**	**Standardized *b****	***t*(195)**
Intercept	4.26	0.549		7.76**
Diagnosis	0.81	0.113		7.19**
Stress	0.20	0.037		5.33**
Gender	–0.47	0.113		–4.21**
IQ	–0.01	0.004		2.10*
(E) Regression Model of Equation (7) [$R^2 = 0.41$]				
IV	**Unstandardized *b***	**Standard Error (*b*)**	**Standardized *b****	***t*(194)**
Intercept	3.84	0.628		6.11**
Diagnosis	0.34	0.364		0.92
Stress	0.19	0.038		4.96**
Gender	–0.48	0.112		–4.27**
IQ	–0.01	0.005		–2.51*
Problems	0.02	0.018		1.37

Note: * $p < .05$, ** $p < .01$. Standardized regression coefficients b^* are not shown when there is binary IV.

ruling out the effects of potential confounding variables, the choice of background covariates demands very careful consideration. In our example, the student appropriately includes participant *Gender* and *IQ* as covariates in her study of the relationship between prior *Diagnosis*, *Stress*, and *Poor Workplace Functioning*. The student estimates a final planned regression equation that contains all four IVs to address this issue.

$$\hat{Y} = b_0 + b_1 Diagnosis + b_2 Stress + b_3 Gender + b_4 IQ \quad (4)$$

The results are presented in Table 13.1D. In Equation (4), $b_0 = 4.26$ is the predicted value of

Poor Workplace Functioning when *Diagnosis* = 0 (comparison group), *Stress* = 0, *Gender* = 0 (female), and *IQ* = 0. No one in the sample could have an *IQ* = 0, so the intercept will *not* represent an interpretable value. Later in the context of nonlinear effects, we will show how to center the IVs so that estimates of all parameters including the intercept can be interpreted. Considering first the two binary IVs, $b_1 = 0.81$ represents the difference in *Poor Workplace Functioning* of the diagnostic minus the comparison groups and $b_3 = -0.47$ represents the difference in *Poor Workplace Functioning* of males minus females, both holding constant the value of each of the other IVs in the equation. Since *Gender* is coded male = 1, female = 0, males have better functioning in this particular workplace setting (lower mean on DV), holding the other IVs constant. Turning now to the IVs corresponding to rating scales, $b_2 = 0.20$ represents the increase in *Poor Workplace Functioning* corresponding to a 1-unit increase in *Stress* and $b_4 = -0.01$ represents the change in *Poor Workplace Functioning* corresponding to a 1-point increase in *IQ*, both effects holding each of the other variables in the equation constant.

As noted earlier, $R^2 = 0.41$ represents the proportion of variation accounted for in the DV by the set of four IVs in the regression equation. This statistic can be employed to characterize the contribution of sets of variables to prediction over and above the other predictors in the equation. By comparing the $R^2 = 0.41$ of Equation (4), which contains *Diagnosis*, *Stress*, *Gender*, and *IQ* as predictors, with the $R^2 = 0.34$ of Equation (3), the independent contribution of the background variables *Gender* and *IQ* (potential confounders) can be evaluated. Here the potential confounders of *Gender* and *IQ* account for $0.41 - 0.34 = R^2_{\text{gain}} = 0.07$ or 7 percent of the variation in the DV of *Poor Workplace Functioning* above and beyond the two variables of theoretical interest, *Diagnosis* and *Stress*. Often of more interest is the comparison of the results of Equation (4) with Equation (5) below – the researcher can evaluate the independent contribution of the two variables of theoretical interest, over and above the background variables.

$$\hat{Y} = b_0 + b_3 Gender + b_4 IQ \tag{5}$$

In each case, the set of variables to be tested are deleted from the full equation and the reduced regression equation is estimated. In the present case, reduced regression equation (5) with only the background variables (not shown in Table 13.1) yields an $R^2 = 0.18$, whereas the full regression equation (4) yields an $R^2 = 0.41$. Thus, the gain in prediction from the set of the two variables of theoretical interest over the prediction from the background variables only is $0.41 - 0.18 = R^2_{\text{gain}} = 0.23$. This gain in prediction is the square of the semipartial correlation, a simple standardized measure of the effect size of adding the set of variables. Put another way, the squared semipartial correlation is the proportion of increase in accuracy of prediction of the DV by the addition of one or more IVs to a regression equation. Cohen (1988) provided descriptive norms for small (0.02), moderate (0.13), and large (0.26) squared multiple correlations, which we use to interpret the squared semipartial correlation. Thus, 0.23 represents slightly less than a large effect.

The statistical significance of the gain in prediction can be tested:

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{reduced}})/m}{(1 - R^2_{\text{full}})/(n - k - m - 1)} = \frac{(0.41 - 0.18)/2}{(1 - 0.41)/195} = 37.7, \tag{6}$$

with $df = (m, n - k - m - 1) = (2, 195)$. In the equation, k refers to the number of predictors in the base or control set and m to the number of predictors whose contribution above the base set is being assessed. Here we have $k = 2$ control variables (*IQ*, *Gender*) plus $m = 2$ added variables (*Diagnosis*, *Stress*) for a total of $(k + m) = 4$ IVs – the four IVs represented in Equation (4). In Equation (6), R^2_{full} is the R^2 from the full regression equation with all of the predictors and R^2_{reduced} is the R^2 from the reduced regression equation (here, represented in equation 5 with $k = 2$ control predictors). In our example, the gain in prediction, R^2_{gain} from the addition of *Stress* and *Diagnosis* to an equation already containing IQ and *Gender*, is statistically significant, $F(2, 195) = 37.7$, $p < .001$. The gain in prediction formula is very useful in assessing effects of confounding variables in clinical research. This procedure is implemented in standard statistical packages (e.g., SAS, SPSS).

(E) *Adding a Redundant Predictor*. Suppose one of the members of the student's Ph.D. committee has developed a 101-point measure of current psychological problems (*Problems*) that he gives to all graduating seniors in the college. He strongly encourages the Ph.D. student to add this IV *Problems* to the regression analysis, stating that "the results would be interesting." The student runs the new regression, Equation (7).

$$\hat{Y} = b_0 + b_1 Diagnosis + b_2 Stress + b_3 Gender + b_4 IQ + b_5 \text{Pr}oblems \quad (7)$$

The results of this analysis are presented in Table 13.1E. As can be seen, the effect of *Diagnosis* is no longer statistically significant, $b_1 = 0.34$, $t(194) = 0.92$, *ns*. The redundant predictor overlaps greatly with *Diagnosis* – the ability of *Diagnosis* to make a unique contribution to the prediction of the DV is greatly reduced when *Problems* is included in the regression equation. Note the large increase in the standard error of the test of b_1 from 0.113 in Equation (4) (Table 13.1D) to 0.364 in Equation (7) (Table 13.1E). In addition, exactly how one interprets the effect of prior diagnosis controlling for a measure of more recent psychological problems is not clear.

Great care must be taken when selecting IVs to enter into a regression equation. Choosing IVs that are measures of similar constructs (e.g., Beck and Hamilton measures of depression) or that represent the same IV at different time points (e.g., BDI at age 18 and age 19) can lead to this problem. This problem of redundancy in predictors can be diagnosed by using a measure known as the variance inflation factor (*VIF*), which indicates how much the squared standard error (SE^2) of a regression coefficient for a specific predictor is multiplied due to the correlation of that predictor of interest with the set of other predictors in the model. Note that *VIF* = 1 if there is no correlation of the predictor of interest with any other predictors in the equation. The *VIF* grows larger above 1 as the correlation of the predictor of interest with the other predictors increases. *VIF*s > 10 indicate seriously problematic redundancy of the predictor of interest with the other IVs. In Equation (4), the *VIF* for *Diagnosis* was 1.05; in Equation (7), it was 10.99, no longer within the acceptable range. Cohen and colleagues (2003, Chapter 10) discuss this problem of predictor redundancy, known as multicollinearity, and offer solutions. Many of the solutions involve reducing the redundancy of the IVs, for example by dropping redundant IVs from the regression model or forming a single composite variable that represents the set of redundant IVs. Giving careful thought to what IVs should be included in the model and including only background variables that are conceptually *not* part of the theoretical construct of interest can help prevent this problem. Investigators who throw additional IVs into their model to see what happens run the risk of obscuring important findings, as in the regression analysis reported in Table 13.1E.

Categorical IVs

IVs that comprise multiple distinct groups such as diagnostic groups or treatment conditions require special consideration in regression analysis. Consider a randomized experiment in which there are three treatment conditions for depression: (a) cognitive behavioral therapy (Therapy), (b) antidepressant medication (Drug), and (c) Control. The DV is a measure of depression given after the completion of therapy. The three treatments represent different categories; they do *not* represent a numerical scale. Two different coding schemes are typically useful in clinical research: dummy codes and contrast codes. The coding scheme chosen should be the one that best represents the researcher's hypotheses.

Dummy Codes. With dummy codes, one group is designated as the reference group against which each of the other groups is compared. The group designated as the reference group is chosen because it provides a useful comparison (e.g., a control group; standard therapy group). Each remaining group is each represented by a separate dummy variable. The non-reference groups are represented by $G - 1$ dummy variables, where G is the total number of groups.[4] Table 13.2A, dummy coding scheme 1,

Table 13.2 Illustration of Coding Schemes

A. Two Dummy Variable Coding Schemes

1. Control as Reference Group

Group	D_1	D_2
Therapy	1	0
Drug	0	1
Control	0	0

2. Therapy as Reference Group

Group	D_1	D_2
Therapy	0	0
Drug	1	0
Control	0	1

B. Contrast Coding Scheme

Group	C_1	C_2
Therapy	+0.33	+0.5
Drug	+0.33	–0.5
Control	–0.67	0

displays a dummy variable scheme with the control group as the reference group.

Equation (8) represents the comparison of the three treatment groups:

$$\hat{Y} = b_0 + b_1 D_1 + b_2 D_2 \quad (8)$$

In Equation (8), using dummy variable scheme 1, b_0 is the mean of the Control group, b_1 is the mean of the Therapy group minus the mean of the Control group, and b_2 is the mean of the Drug group minus the mean of the Control group. D_1 and D_2 represent the dummy codes. No comparison is directly possible between the Therapy and Drug groups using dummy variable scheme 1. Taking another perspective, many clinicians regard cognitive-behavioral therapy as the "standard therapy" for depression. If we were to take the Therapy group as our reference, then we get dummy variable coding scheme 2 in Table 13.2A. With this coding scheme, b_0 is the mean of the Therapy group, b_1 is the mean of the Drug group minus the mean of the Therapy group, and b_2 is the mean of the Control group minus the mean of the Therapy group. The workplace problems example presented in Table 13.1 illustrated the use of dummy variable coding when there are two groups: Gender (1= Male, 0 = Female) and Diagnosis (1 = Diagnosis; 0 = No Diagnosis).

Contrast Codes. Contrast codes permit comparisons involving two or more group means. For example, a researcher might hypothesize that (a) the two active therapy groups (Therapy, Drug) would differ from the Control, but that (b) the two groups would not themselves differ. Table 13.2B presents a contrast coding scheme that represents these hypotheses and Equation (9) represents the regression equation.

$$\hat{Y} = b_0 + b_1 C_1 + b_2 C_2 \quad (9)$$

Equation (9) repeats Equation (8) replacing the dummy codes with contrast codes. b_0 is now $\bar{M}$ = the unweighted mean of the three treatment groups, $\bar{M} = (M_{Therapy} + M_{Drug} + M_{Control})/3$. b_1 is the difference between the unweighted mean of the two active treatment groups, $(M_{Therapy} + M_{Drug})/2$ minus the mean of the Control group, $M_{Control}$. b_2 is the difference between the mean of the therapy group $M_{Therapy}$ minus the mean of the drug group M_{Drug}. When only two groups are involved (e.g., Treatment (T), Control (C)), they would be coded as $T = +0.5$ and $C = -0.5$.

Once again, we emphasize that researchers should choose the coding scheme that best represents their hypotheses. In writing the results, they should also clearly report the coding scheme and the specific values that were used for each group. In some articles we reviewed in *JAP* and *JCCP*, the coding scheme was not clearly specified, so the numerical results could not be interpreted. With coding schemes for group variables, standardized solutions are not normally reported. As noted previously, the standardized solution is very sensitive to the proportion of participants in each group, whereas the unstandardized solution is not. Cohen and colleagues (2003, Chapter 8) provide a discussion of coding schemes for categorical variables, including less frequently used schemes that are not considered here. As we will see later in this chapter, a clear understanding of the coding schemes for categorical IVs can facilitate the interpretation of the results of more complex regression models.

Assumptions of Linear Regression: Diagnosing Violations

Linear regression makes several assumptions. When the assumptions are not met, the estimates of the regression coefficients, their standard errors, or both may be biased. Bias means that the estimate based on the sample will not on average be equal to the corresponding population value. The implication is that estimates of the regression coefficients, R^2, hypothesis tests, and confidence intervals may all be incorrect. Here we briefly consider the major assumptions and how to assess the extent to which each assumption is violated. Assessing violations of some assumptions involves careful attention to plots of the residuals, $e = Y - \hat{Y}$, which can highlight problems of regression models. In the second half of the chapter, we consider more advanced regression models that can address violations of some of these assumptions.

1. *Correct Specification of the Relationship between IVs and the DV*. The models we have considered so far all assume that there is a linear (straight-line) relationship between each of the IVs and the DV. In fact, nonlinear relationships are possible. These can be diagnosed by plotting the residuals against the predicted values of Y (i.e., $\hat{Y}$). Figure 13.2A portrays a correctly specified linear relationship; Figure 13.2B portrays a nonlinear relationship between X and Y.

2. *No Measurement Error in the Independent Variable*. In regression equations with one predictor, measurement error in the IV leads to an estimate of the regression coefficient that is attenuated, closer to 0 than it should be. With

more than one IV, regression coefficients will typically be attenuated, but individual regression coefficients can occasionally be too large in magnitude. Measurement error is detected by the examination of the reliability of each of the IVs. Most reliability coefficients can vary from 0 to 1; as the value gets closer to 1, attenuation diminishes.

3. *Constant Variance of Residuals (Homoscedasticity).* The variance of the residuals around the regression line is assumed to be constant for any value of $\hat{Y}$. When this assumption is violated, the regression coefficients remain unbiased, but significance tests and confidence intervals will be incorrect. In addition, the interpretation of R^2 is compromised because the variation now depends on the predicted value of Y. Again, the plot of the residual versus $\hat{Y}$ can be informative (see Fig. 13.2C).

4. *Nonindependence of the Residuals*. When data are collected in groups (clustering) or from the same individuals over time (autocorrelation), the values of the residuals may not be independent. Statistical tests for detecting clustering (the intraclass correlation coefficient) and serial dependency (Durbin-Watson test) exist. Typically far more informative is a careful consideration of the structure of the data. Are there substantive reasons to believe that the residuals might not be independent? Nonindependence of the residuals does not bias the regression coefficients but leads to incorrect significance tests and confidence intervals.

5. *Normal Distribution of Residuals*. The residuals are assumed to have a normal distribution. Although minor violations of this assumption do not have material effects on the results, more severe violations can lead to incorrect significance tests and confidence intervals. Plots of the residuals (e.g., histograms) can help the researcher visualize the shape of

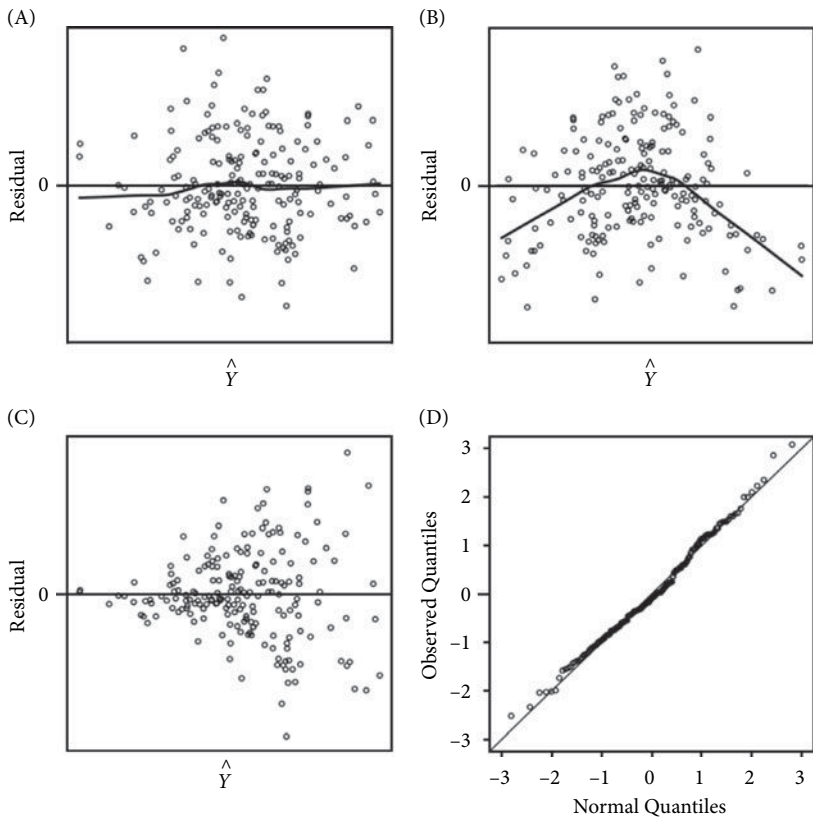


Figure 13.2 Residual plots assessing assumptions in MR. (**A**) *Linearity*. Residuals are plotted against the predicted values of Y ($\hat{Y}$). For this dataset, the nonparametric LOWESS line closely parallels the 0-line, indicating the linearity assumption is met. (**B**) *Nonlinearity*. For this dataset, the LOWESS line does not parallel the 0-line, indicating that curvilinearity is present in the data. (**C**) *Heteoscadasticity*. For this dataset, the spread of the residuals around the 0-line increases with increasing $\hat{Y}$, indicating that the assumption of constant variance of residuals (homoscedasticity) is violated. (**D**) *Normality*. A q-q plot of the residuals against the expected quantiles of a normal distribution closely parallels a straight line, indicating that normality of the residuals is met.

the distribution. Most informative is a q-q plot against a normal distribution (see Fig. 13.2D). To the extent that the distribution of the residuals follows a normal distribution, this graph will appear as a straight line, making violations easy to detect.

Models with Curvilinear Effects and Interactions

Regression analyses in clinical psychology often assume that relationships have a linear form, but two general forms of nonlinear relationships are hypothesized and tested with some regularity. First, the relationship between the IV and DV may be hypothesized to have a curvilinear relationship. Seven percent (4/59) of the articles reviewed in *JAP* and 13 percent (6/46) of the articles reviewed in *JCCP* considered a curvilinear relationship, most often in the form of growth trajectories (considered later in this chapter). Second, relationships may be hypothesized to involve interactions in which two (or more) IVs combine to predict the DV. Sixty-four percent (38/59) of the articles reviewed in *JAP* and 61 percent (28/46) of the articles reviewed in *JCCP* tested at least one interaction.

Curvilinear Models

Curvilinear relationships are most often represented as quadratic models in which curvilinearity is represented by a curve with a single bend. As illustrated in Figure 13.3, quadratic models are quite versatile – they are able to represent many (but not all) of the forms of curvilinear relationship between an IV and DV. Consider the relationship between the level of anxiety (*Anx*) of a student before giving a talk and the student's subsequent performance (*Perf*). At extremely low anxiety, the student may not be fully engaged, and performance is poor. As anxiety increases, performance increases up to a maximum. However, as anxiety continues to increase, performance now falls off. The relationship between *Anx* and *Perf*, depicted in Figure 13.3A, is that of an upside-down U – an inverted-U–shaped function. This relationship is described by Equation (10):

$$\widehat{Perf} = b_0 + b_1\, Anx + b_2\, Anx^2 \qquad (10)$$

The added term, the square of the anxiety score Anx^2, carries the curvilinear aspect of the relationship of *Anx* to *Perf*. The test of significance of the b_2 coefficient in Equation (10) informs us whether the quadratic term adds to prediction

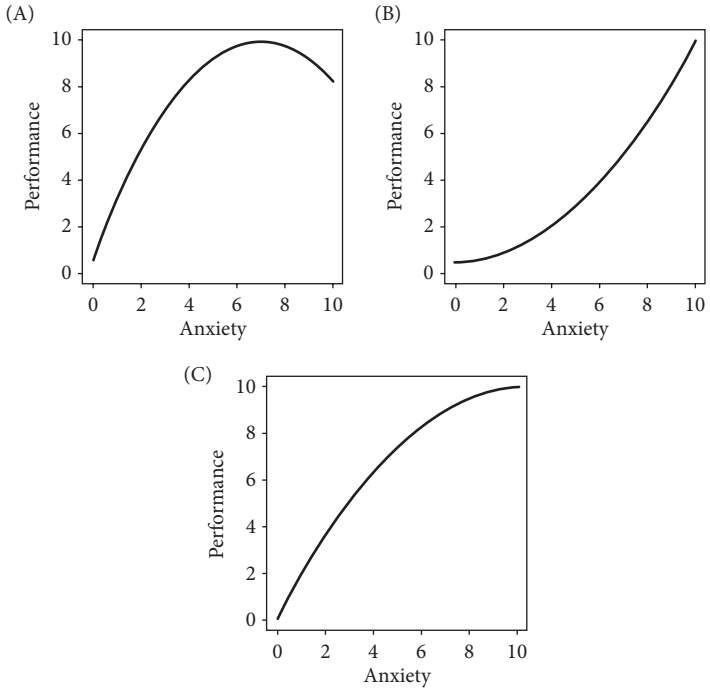


Figure 13.3 Illustrations of some curvilinear relationships that may be represented in quadratic regression models. (**A**) An inverted-U–shaped relationship (described in text). (**B**) An accelerating relationship (increasing slope with increasing values of *X*). (**C**) A decelerating relationship (decreasing slope with increasing values of *X*).

over and above the linear term. A negative b_2 coefficient indicates an inverted-U–shaped relationship ("frown-shaped" ☹); a positive b_2 coefficient indicates a U-shaped relationship ("smile-shaped" ☺). Figures 13.3B and 13.3C present illustrations of two of the forms of relationships that can be represented with the quadratic regression model presented in Equation (10). Interpretation of the b_0 and b_1 coefficients requires a foray into an important strategy in MR, centering predictors.

Centering Predictors in Regression Equations with Higher-Order Terms. Terms such as X^2 (or X^3 or XZ) that are higher-order functions of the original IVs are called *higher-order terms*. Higher-order terms add complexity to the interpretation of all of the lower-order terms that the higher-order term comprises. In Equation (10) Anx^2 is the higher-order term and Anx and the intercept b_0 are lower-order terms. In regression equations with higher-order terms, all lower-order terms are conditional; they represent the relationship of the lower-order predictor to the criterion *only* at the value of 0. Figure 13.4A shows a quadratic relationship (inverted-U) of a predictor X to Y. The predictor X ranges from 1 to 6 – the value 0 is *not* represented on the scale. The slope of the relationship of X to Y is different for each value of X. $b_1 = 4.751$ represents the slope of the regression of Y on X at $X = 0$ only, highlighted by a vertical arrow in the figure. The heavy tangent line touching the curved regression equation at $X = 0$ depicts this slope. Recall X is on a 1-to-6 scale, so $X = 0$ is a value that cannot exist in the data! b_1 reflects the hypothetical slope if the relationship of X to Y given in the equation were projected down to $X = 0$, not meaningful information.

To remedy this problem, we subtract the mean of X from each value on the predictor to create X_C, the centered value of X:

$$X_C = X - \text{MEAN}(X) \tag{11}$$

The mean of X_C will equal 0 since X_C is a deviation score. Thus, the b_1 coefficient now represents the regression of Y on X at the mean of the predictor X. The effect of mean centering is illustrated in Figure 13.4B. Note first that the x axis now runs from –3 to +3, instead of 0 to 6. The vertical arrow beneath Figure 13.4B points to the value of zero on X_C. The heavy tangent line to the curve directly above the arrow represents the regression of Y on X_C at centered $X_C = 0$, or at the arithmetic mean of the distribution of the original predictor X. In the centered regression equation given below Figure 13.4B, the b_1 coefficient for X is now 1.256, which tells us that the criterion Y is still increasing at the mean on the predictor (for the anxiety and performance example, that performance *Perf* is still increasing at the mean of *Anx* in the sample). Note that the b_2 coefficient has not changed; it is still $b_2 = -0.582$. The highest-order term represents the shape of the regression equation, and this does not change when one centers the predictor; nothing is lost or distorted about the shape of the relationship of X to Y by centering predictor X. Centering yields a meaningful

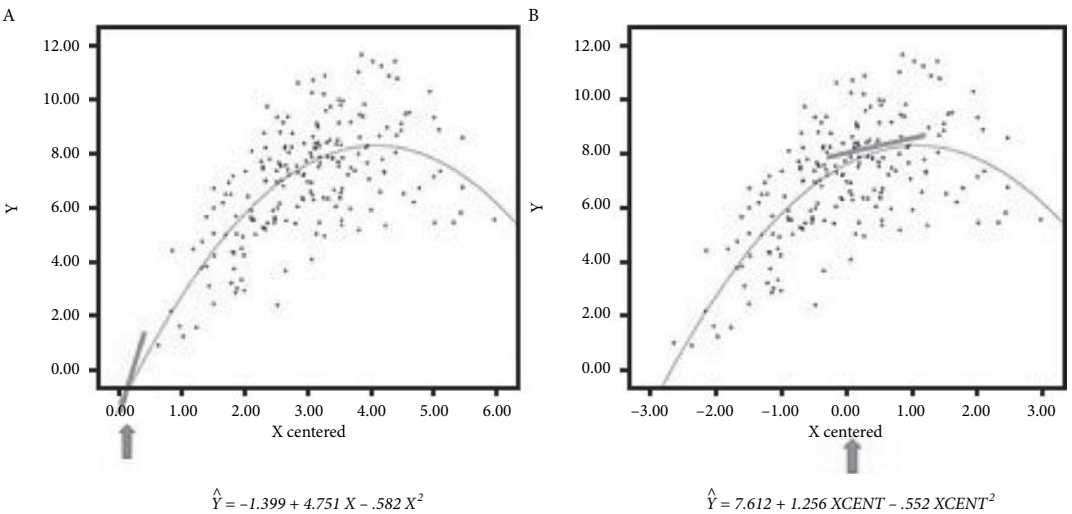


Figure 13.4 The effect of centering the predictor in a quadratic regression equation. (**a**) b_1 is the slope of the tangent to the curve (*heavy line*) at uncentered $X = 0$. (**b**) b_1 is the slope of the tangent to the curve at $X_c = 0$, the mean of centered X. Note that the shape of the curves is identical; only the scaling of the x axis has changed.

interpretation of the lower-order coefficients, the regression of Y on X at the arithmetic mean of X. It also yields a second useful interpretation of b_1 as the average regression of Y on X across the whole range of X. Finally, what happens to the intercept b_0? Recall that b_0 is the value of Y when all predictors in the equation equal zero. Thus, b_0 is a conditional coefficient that changes when the predictor X is centered. In Figure 13.4A, in which X has *not* been centered, $b_0 = -1.399$ (note the minus sign), the value of Y that is predicted if X were equal to zero; there are in reality no negative values of the criterion, so the value –1.399 comes from projecting the regression curve down to zero on X and to –1.399 on Y. In Figure 13.4B, $b_0 = 7.612$, indicating that at the arithmetic mean anxiety of the sample (X_C), the predicted value of Y is 7.612. This is easily seen by looking for the value of Y on the regression curve at $X_C = 0$.

The maximum (or minimum) of the quadratic curve will occur when $X = \frac{-b_1}{2b_2}$. When $X = \frac{-b_1}{2b_2}$ occurs outside the observed range of the IV, the curve will not reach a maximum or minimum value within the range of the data, permitting many forms of curvilinear relationship to be represented, as shown in Figure 13.3.

Other strategies of representing curvilinear relationships are sometimes used. When the relationship between X and Y is only increasing or only decreasing, X, Y, or both can be transformed to attempt to represent the relationship as a straight line. A second strategy is to include higher-order polynomial terms in the regression equation so that the function can represent more than one bend (e.g., $\hat{Y} = b_0 + b_1X + b_2X^2 + b_3X^3$). A variety of nonparametric methods (e.g., LOWESS smoother) exist for representing the X–Y relationship based solely on the available data. Further, there are nonlinear regression models that use functions other than polynomial terms. Cohen and colleagues (2003, Section 4.2.2 and Chapter 6) and Cook and Weisberg (1999) consider these methods in more detail.

Interaction Models

Continuous × Continuous Variable Interactions. IVs may interact, yielding a combined effect that is different from the sum of their first-order effects. In Equation (12), we have two continuous predictors X and Z and their interaction XZ:

$$\hat{Y} = b_0 + b_1X + b_2Z + b_3XZ \qquad (12)$$

The first-order effects are the linear relationships represented in the terms $[b_1X + b_2Z]$. The new term b_3XZ represents the interaction between predictors X and Z that contributes to the prediction of Y over and above the sum of the linear relationships. The interaction term is calculated by multiplying the score on X by the score on Z for each case.

In psychology we often describe interactions as moderator effects. There is a focal variable X and a second variable Z that moderates the impact of the first variable on the outcome; the second variable Z is referred to as the *moderator*. To illustrate, we developed a hypothetical example based loosely on a study by La Greca, Silverman, Lai, and Jaccard (2010). In our simulated example ($n = 415$), perceived *Threat* (X) predicts *Symptoms* of distress (Y), but the positive relationship of *Threat* to *Symptoms* becomes weaker as the level of the moderator social *Support* (Z) increases. The relationship is depicted in Figure 13.5. The three regression lines in Figure 13.5 represent three values of *Support* chosen across the range of reported social support. When *Support* is low (the top regression line), the relationship of *Threat* to *Symptoms* is strongly positive; at the mean level of *Support* (the middle line), the relationship is still positive but weaker. Finally, when *Support* is high (the bottom line), the positive relationship of *Threat* to *Symptoms* is quite weak. In addition, there is a first-order effect: on average as *Support* increases, the average level of *Symptoms* decreases (i.e., the average elevation of the regression lines decreases as *Support* increases).

In Equation (12), the interaction XZ term is a higher-order term; thus, as in the case of the quadratic equation, the b_1 and b_2 coefficients are conditional. The b_1 coefficient is the regression of Y on

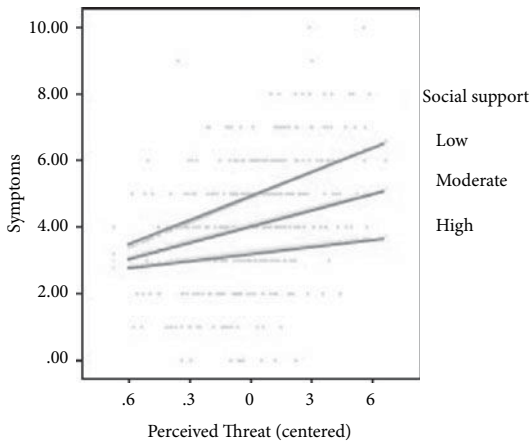


Figure 13.5 A buffering interaction: psychological *Symptoms* following perceived *Threat* at three values of social *Support*. Both the overall level of *Symptoms* and the strength of relationship of *Threat* to *Symptoms* decrease with increasing *Support*.

X at $Z = 0$; the b_2 coefficient is the regression of Y on Z at $X = 0$. Thus, following the logic of the discussion of centering predictors in equations with higher-order terms, we center both X and Z before forming the interaction term. The analysis is carried out with the two centered predictors (X_C, Z_C) and the interaction term $X_C Z_C$ formed by multiplying together the centered predictors:

$$X_C = X - \text{MEAN}(X), \tag{13a}$$

$$Z_C = Z - \text{MEAN}(Z), \text{ and} \tag{13b}$$

$$X_C Z_C = X_C \times Z_C. \tag{13c}$$

Note that we do not typically center the dependent variable *Symptoms*; the predicted scores $\hat{Y}$ are kept in their original scale.

For the data depicted in Figure 13.5, the centered regression equation containing the interaction is as follows:

$$\widehat{Symptoms} = 4.083 + 0.162\ Threat_C - 0.616\ Support_C - 0.068\ \text{Stress}_C \times Support_C \tag{14}$$

If the interaction term $b_3 = -0.068$ is significantly different from zero, this indicates that there is an interaction between the two predictors. To find an effect size for the interaction, we use the gain in prediction approach described in an earlier section. We estimate the regression equation deleting the interaction. For our example,

$$\widehat{Symptoms} = 4.190 + 0.165\ Threat_C - 0.619\ Support_C. \tag{15}$$

For the full interaction model, $R^2_{\text{full}} = 0.352$; for the equation containing only linear terms, $R^2_{\text{linear}} = 0.337$. $R^2_{\text{full}} - R^2_{\text{linear}} = 0.015$ or 1.5 percent. This result appears at first glance like a tiny gain in prediction; we note that many interactions are very small in the incremental proportion of variation accounted for (Chaplin, 1991). Yet these interactions can have dramatic effects on outcomes, as we will see in what follows.

Because the predictors are centered, the conditional b_1 and b_2 coefficients have straightforward interpretations. The $b_1 = 0.162$ coefficient for $Threat_C$ indicates a positive regression of *Symptoms* on $Threat_C$ at the sample arithmetic mean level of $Support_C$. The $b_2 = -0.616$ coefficient for $Support_C$ indicates a negative regression of *Symptoms* on $Support_C$ at the sample arithmetic mean of *Threat.* Further, b_1 and b_2 can be interpreted as the average relationship of $Threat_C$ and $Support_C$, respectively, to *Symptoms*.

More insight can be gained on the conditional relationship of $Threat_C$ to *Symptoms* as a function of $Support_C$. We rearrange regression equation (12) into a *simple regression equation* that shows the prediction of Y from focal predictor X as a function of the moderator Z:

$$\hat{Y} = (b_0 + b_2 Z_C) + (b_1 + b_3 Z_C)\ X_C \tag{16}$$

The term $(b_1 + b_3 Z_C)$ is the *simple regression coefficient* for predictor X_C; the term shows that the prediction of Y from X_C changes as the value of Z_C changes – this is the moderator effect of Z_C on the relationship of X_C to Y. The term $(b_0 + b_2 Z_C)$ is the intercept; this term shows how the intercept changes as the value of Z_C changes. The rearranged regression equation for predicting *Symptoms* from $Threat_C$ as a function of $Support_C$ is as follows:

$$\widehat{Symptoms} = (4.083 - 0.616\ Support_C) + (0.162 - 0.068\ Support_C)\ Threat_C. \tag{17}$$

$Support_C$ is a continuous IV, so we may ask at what values of $Support_C$ we might examine the relationship of $Threat_C$ to *Symptoms*. Sometimes well-established cutoff values are defined. For example, a score of 30 or above on the BDI indicates severe depression; $Z = 30$ would be a good choice in this case. When available, established cutoff values should be used. In the absence of established cutoff values, we recommend examining the *simple regression equations* at three values of the moderator: the arithmetic mean of the moderator, one standard deviation above the mean of the moderator, and one standard deviation below the mean of the moderator. These values typically ensure that we are staying within the range of the data in which there are a sufficient number of cases. The three regression lines in Figure 13.5 are these simple regression equations. The light gray dots are the actual data points. Were we to go out two standard deviations beyond the mean of *Support*, we would find almost no cases. This point is critical: one should choose only those values of the moderator Z that yield simple regression equations where there are a reasonable number of cases.

The standard deviation of $Support_C = 1.336$. The mean of $Support_C$ is 0. We substitute –1.336, 0, and 1.336 into Equation (17) to find three simple regression equations. In Equation (18) we show the

substitution for the low, mean, and high values of centered support.

(A) At LOW $Support_C = -1.336$:

$$\widehat{Symptoms} = (4.083 - 0.616 \times -1.336) + (0.162 - 0.068 \times -1.336)\ Threat_C$$

$$\widehat{Symptoms} = 4.906 + 0.253\ Threat_C \quad (18a)$$

(B) At the arithmetic MEAN of $Support_C = 0$:

$$\widehat{Symptoms} = (4.083 - 0.616 \times 0) + (0.162 - 0.068 \times 0)\ Threat_C$$

$$\widehat{Symptoms} = 4.083 + 0.162\ Threat_C \quad (18b)$$

(C) At HIGH $Support_C = 1.336$:

$$\widehat{Symptoms} = (4.083 - 0.616 \times 1.336) + (0.162 - 0.068 \times 1.336)\ Threat_C$$

$$\widehat{Symptoms} = 3.260 + 0.071\ Threat_C. \quad (18c)$$

As $Support_C$ increases from Equation (18a) to (18b) to (18c), the simple regression coefficient for the regression of *Symptoms* on $Threat_C$ decreases from 0.253 ($p < .001$) to 0.162 ($p < .001$) to 0.071 ($p = ns$). The stress-buffering hypothesis predicts that social support is protective (i.e., weakens the impact of threat on negative outcomes).

MacKinnon and colleagues (Chapter 15 in this volume) provide a fuller introduction to moderation and Aiken and West (1991) provide a full account of the analysis of interactions in MR, along with SPSS syntax for carrying out the regression analysis, computing simple regression equations, and testing for significance of the simple regression coefficients. They develop more complex interactions, including changes in curvilinearity as a function of the moderator, and interactions involving more than two predictors.

Continuous × Categorical Interactions. Often we are interested in whether a categorical moderator, for example a previous clinical diagnosis versus its absence, modifies the relationship of an IV to an outcome. For the case of a binary IV, the transition from the continuous variable interaction to continuous × categorical interactions is straightforward. We replace the continuous moderator Z of Equation (12) with a categorical moderator. The categorical moderator is dummy coded or contrast coded, as presented earlier in the chapter. For example, if the moderator were dummy coded:

$$\hat{Y} = b_0 + b_1 X_C + b_2 D + b_3 X_C D \quad (19)$$

As an illustration, we draw on ideas presented by Doron-LaMarca, Vogt, King, King, and Saxe (2010) about the psychological sequelae of physical injury as a function of time since injury. For our simulated example, we consider $n = 120$ professional athletes who have recently experienced a serious injury that at present is keeping them from playing. In addition, all 120 athletes have experienced another injury in the past. Half ($n = 60$) have experienced full recovery from the previous injury, whereas the other half ($n = 60$) have residual effects of the previous injury. Each athlete is interviewed following the recent injury; athletes rate their *Confidence* to recover sufficiently to perform the next season on an 81-point scale, where 0 = no confidence and 80 = highest possible confidence. We predict *Confidence* from *Weeks* that have elapsed between the time of the recent injury and the time of the interview (ranging from 3 to 21 weeks), and *Recovery* from the previous injury using a dummy variable where 1 = Recovered and 0 = Not Recovered. Figure 13.6 illustrates the prediction of *Confidence* from *Weeks* since injury for the two groups (solid vs. dashed regression line for Recovered and Not Recovered, respectively). The regression lines are not parallel across the two groups, indicating there is an interaction between *Recovery* and *Weeks* since injury in predicting *Confidence*. Those who have recovered

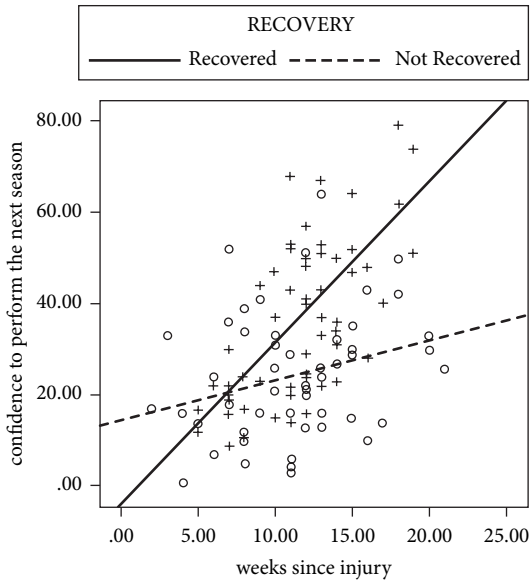


Figure 13.6 A synergistic interaction: *Confidence* of injured athletes to perform in the next season as a function of *Weeks* since injury and *Recovery* from previous serious injury.

from their previous injury have a much steeper slope of the relationship of *Weeks* to *Confidence* than those with residual injury effects. This shows that the injury history coded in the dummy variable *Recovery* moderates the relationship of *Weeks* to *Confidence*. This interaction can be described as synergistic or enhancing: on average, *Confidence* increases as *Weeks* since the injury increase. On average, *Confidence* increases if one has recovered successfully from a previous serious injury. Taken together, *Weeks* and *Recovery* combine to produce higher levels of *Confidence* than would be expected from the additive effects of *Weeks* and *Recovery*.

In Figure 13.6 the data points for the Recovered group are shown by plus signs (+) and those for the Not Recovered group are represented by open circles (○). Weeks since injury is left uncentered; the time from injury to interview is in the original units of number of weeks. Recall from Figure 13.4 portraying a quadratic relationship that centering predictors leaves the form of the relationship of the IVs to the outcome unchanged. The figure would look identical if *Weeks* had been centered. Only the scaling of the *x* axis would change.

Recall that multiple options exist for coding a categorical predictor. Here we use the dummy variable coding scheme (1, 0) for two groups presented earlier. We do not center the dummy-coded categorical variable. *Weeks* is centered at the mean of weeks since injury = 11.48, $Weeks_C$ = *Weeks* – Mean(*Weeks*). We estimate the regression equation with dummy-coded *Recovery*, $Weeks_C$, and their interaction *Recovery* × $Weeks_C$:

$$\widehat{Confidence} = b_0 + b_1 Weeks_C + b_2 Recovery + b_3 Recovery \times Weeks_C$$

$$\widehat{Confidence} = 24.592 + 0.865\, Weeks_C + 11.895\, Recovery + 2.661\, Recovery \times Weeks_C \quad (20)$$

We also present the simple regression equation for each group of 60 athletes.

(A) For the Not Recovered group (*Recovery* = 0)

$$\widehat{Confidence} = 24.592 + 0.865\, Weeks_C \quad (21a)$$

(B) For the Recovered group (*Recovery* = 1)

$$\widehat{Confidence} = 36.487 + 3.526\, Weeks_C \quad (21b)$$

Examining the simple regression equations within each group gives important insight into the meaning of the coefficients in the overall regression, Equation (20). Recall that in regression equations with higher-order terms, the lower-order regression coefficients are conditional, interpreted only at zero (0) on the other predictor. In Equation (17), the regression coefficient $b_1 = 0.865$; this is the regression of *Confidence* on $Weeks_C$ for the Not Recovered group, coded *Recovery* = 0. The intercept b_0 is the value of *Confidence* when *both* predictors equal zero; this occurs at the mean number of $Weeks_C$ (where $Weeks_C = 0$) and *Recovery* = 0 (Not Recovered group). The b_2 coefficient is the difference between predicted *Confidence* of the group coded 1 minus the predicted *Confidence* of the group coded zero = 36.487 – 24.797 = 11.690 at $Weeks_C = 0$. Finally, $b_3 = 2.661$ is the regression coefficient for the interaction, the difference between the slope of the *Confidence* group coded 1 (Recovered) minus the slope of the *Confidence* group coded 0 (Not Recovered). Because the regressions in the two groups are linear, this difference in slopes remains constant across the range of $Weeks_C$. Readers are advised that if they publish equations like Equation (20), they must carefully interpret each coefficient; many mistakes in interpretation can be easily made by those who lack understanding of coding and conditional effects in equations with higher-order terms.

There are alternative coding schemes for categorical IVs. In contrast coding, the code variable is centered so the values would be +0.5 for the Recovered group and –0.5 for the Not Recovered group. The b_2 coefficient for *Recovery* and the b_3 coefficient for the interaction will be identical to those we obtained using the dummy coding scheme. In contrast, the b_1 coefficient becomes the average relationship of $Weeks_C$ to *Confidence* for the whole sample taken together (ignoring group membership); the b_0 coefficient becomes the value of *Confidence* at the arithmetic mean number of weeks for the full sample. West, Aiken, and Krull (1996) present a full discussion of the use and interpretation of different coding schemes with continuous × categorical variable interactions.

Missing Data

Missing data are ubiquitous in studies of both community and clinical populations. Participants refuse to answer questions or skip items on questionnaires. They sometimes give inconsistent answers across items (e.g., Question 1. How many alcoholic drinks have you had in your lifetime? Answer: 0 drinks; Question 10. How many alcoholic drinks have you had in the past week? Answer: 4 drinks). Participants in longitudinal studies move, drop out, or are unavailable for every measurement. The best strategy for dealing with missing data is to

take steps to prevent its occurrence (see Ribisl et al., 1996, for methods of minimizing missing data) or to collect the missing information from the participant in another measurement session, from records, or from knowledgeable informants. However, even given the conscientious use of these procedures, missing data will occur. Regression analysis requires a complete dataset on all IVs and the DV.

A number of ad hoc methods of addressing missing data have traditionally been used to create a "complete" dataset for analysis in MR. Listwise deletion uses only participants who have complete data on all variables, but these participants may not be representative of the full sample. Pairwise deletion "tricks" the program into thinking complete data are available. It calculates the means and variances from all participants for whom data are observed on each particular variable and correlations from all participants who have data on each particular pair of variables. The analysis then uses this summary information as if it represented complete data to estimate the regression coefficients.[5] Pairwise deletion does not keep track of which participants are contributing data to each mean and correlation; these sets of participants may differ. Mean imputation simply calculates the mean on each variable for the available cases and replaces the missing data with these means. Each of these methods can potentially lead to problems – biased (incorrect) estimates of regression coefficients, incorrect significance tests and confidence intervals, and decreased power of the statistical test to detect effects when in fact they do exist (see Chapter 12 in this volume). Modern statistical procedures must be used to minimize the impact of missing data on the results of the regression analysis.[6] Our survey of articles published in *JAP* and *JCCP* showed that procedures to address missing data were typically not reported or that nonoptimal traditional methods were utilized. Only 17 (29 percent) of the studies in *JAP* and 20 (43 percent) of the studies in *JCCP* reported the use of a modern procedure for addressing missing data.

Modern approaches to missing data begin with Rubin's (1976) consideration of the potential types of missing data and their effects. In practice, the type of missing data will be unknown and combinations of types of missing data may occur (Enders, 2010). Data can *not* be inspected and the type of missingness determined.[7] The terminology for the different types of missing data can initially be confusing, but this terminology has now become standard.

The simplest (but least common) form of missing data is termed *missing completely at random* (MCAR). MCAR means that missingness – whether a variable is observed or not for each individual participant – is purely random; it is not related to any characteristic of the participant. MCAR might occur if equipment used in an experiment occasionally failed – for example, from electrical power problems. In contrast, data that are *missing at random* (MAR) have a systematic basis, but one that is entirely accounted for by variables measured in the dataset. After these measured variables are statistically controlled, there are no remaining systematic sources of missingness – all remaining sources are random. For example, suppose Spanish-speaking patients in a community study do not complete one particular questionnaire in a battery because it has not been translated into Spanish. The data are MAR: correction for participant language, the only systematic source of missingness, can yield a properly adjusted estimate of the regression coefficient. Finally, data are *missing not at random* (MNAR) if missingness depends on the *unobserved* level of the missing variable. For example, if some bipolar patients fail to complete a symptom report because their symptoms are too severe and they cannot function, data are MNAR. Estimates of the regression coefficients will be biased and might not be correctable. To the extent that variables exist in the dataset (e.g., baseline symptoms measure, clinical psychologist's evaluation) that are related to the participants' reported or unreported score on current symptoms, the bias due to missing data in the estimate of the regression coefficient can be reduced, sometimes substantially. *Auxiliary variables*, variables that predict both missingness and the measures of interest, can greatly reduce bias due to missing data (Collins, Schafer, & Kam, 2001). In essence, inclusion of good auxiliary variables makes MNAR data more like MAR data. Auxiliary variables can lead to proper adjustment of the regression coefficients, even if the auxiliary variables are not themselves of theoretical interest. For example, distance from home to the treatment site might serve as a good auxiliary variable in studies in which participants are dependent upon their own transportation.

Two modern missing data techniques that provide proper adjustment of MAR data are full information maximum likelihood (FIML) estimation and multiple imputation[8] (MI; Enders, 2010; Little & Rubin, 2002; Chapter 19 in this volume). Normally, MR uses an estimation procedure known as ordinary least squares (OLS) to calculate the regression coefficients. OLS requires complete data on all IVs and the DV to estimate the regression

coefficients. In contrast, FIML uses an alternative estimation procedure that directly uses information from all available data, including cases with partially missing data, to provide the estimates. The FIML procedure keeps track of the information contributed by each case. FIML is implemented in several statistical packages (e.g., Mplus) and is easy to use. The primary disadvantages of FIML are that it can be difficult to use auxiliary variables and that it is not optimal when there are nonlinear or interactive relationships in the data.

The second modern procedure, MI, involves a three-step process. In the imputation step (step 1), MI produces m complete copies of the dataset in which the observed values are maintained and missing values are imputed from the available data. In very large datasets only a few copies (e.g., $m = 5$–10) are needed; in clinical datasets in which the median sample size is roughly 200 participants and smaller sample sizes are not uncommon, $m = 20$ to 50 copies will often produce better results in terms of the statistical power of hypothesis tests (see Graham, 2009). A sophisticated version of regression analysis is used to fill in the missing values. The key feature of MI is that it retains the prediction error observed in the original data. In each copy of the dataset different random residual values comparable to those observed in the original dataset are added to each predicted value $\hat{Y}$. Each missing value is replaced by a different $\hat{Y} + e$. This is illustrated in Table 13.3. An advantage of MI is that during this imputation step many auxiliary variables and terms representing quadratic (X^2) and interactive (XZ) effects may be used in the prediction of the missing values (Allison, 2001).

In the analysis step (step 2), identical regression analyses testing the model of interest (e.g., Equation (4)) are performed on each of the m copies (e.g., 20) of the dataset, typically using OLS regression. The results of each analysis are saved. This feature of MI, that auxiliary variables can be easily included in the imputation step but excluded from the analysis step, permits tests of the exact hypotheses of interest.

Finally, in the pooling step (step 3), the results of the m analyses of the dataset are combined. The final estimates of the regression coefficients are simply the means of the corresponding regression coefficients of the m regression analyses. For example, if 20 imputed copies of the dataset were created, the regression coefficient b_1 would be computed as $\bar{b}_1 = \frac{1}{20}\sum_{i=1}^{i=20} b_{1(i)}$, the mean value of b_1, where i is the copy of the imputed dataset. Calculation of standard errors is more complex as it involves computing a weighted combination of the average variability of the standard errors within each copy of the imputed

Table 13.3 Illustration of Multiple Imputation

Case	Original Data		Imputation 1	Imputation 2	…	Imputation 10
	Stress	**Poor Workplace Functioning**	**Poor Workplace Functioning**	**Poor Workplace Functioning**		**Poor Workplace Functioning**
1	3	4	4	4		4
2	4	3	3	3		3
3	4	7	7	7		7
4	4	5	5	5		5
5	5	4	4	4		4
6	5	3	3	3		3
7	6	M	4.83	5.26		4.38
8	6	M	6.07	5.19		2.99
9	7	M	4.05	4.78		6.63
10	8	M	5.50	6.69		4.60

Note: The first two columns present the original data ordered from low to high on the *Stress* variable. Cases 7–0 have missing values on *Poor Workplace Functioning*. Imputations 1, 2, …, and 10 illustrate different copies of the data for *Poor Workplace Functioning* produced by multiple imputation. The observed values in the original data set are preserved; different plausible values identified by a box are imputed on *Poor Workplace Functioning* based on the case's observed value on *Stress*.

dataset and the variability of these standard errors across the *m* imputations. The degrees of freedom for the significance test are also calculated using a complex weighting procedure that reflects the number of observed cases on each variable, the proportion of missing data, the number of imputations, and the correlations between pairs of variables in the dataset. The algorithms can produce fractional degrees of freedom (e.g., $df = 182.7$), which initially seem strange because they cannot occur with complete datasets. The regression coefficient can be tested using the standard *t* test – for example for regression coefficient b_1, $t(df) = \frac{b_1}{SE_{b_1}}$. The MI procedure is easy to use and is implemented in standard statistical packages, including SAS and SPSS, as well as several free-ware packages (e.g., NORM; Schafer, 1999). The results of step 1 require some checking to determine if imputation has been successfully performed; graphical procedures for checking multiple imputation are described in Enders (2010).

The MI and FIML procedures described above do not completely reduce bias in tests of regression coefficients if the data are MNAR. The use of auxiliary variables can reduce bias to the extent they account for missingness. However, there is no guarantee that bias will be eliminated if data are MNAR. Consequently, there is interest in missing-data approaches that adjust for MNAR data. Such approaches for MNAR data have been developed in the statistics literature (see Enders, 2010, 2011, for accessible reviews); indeed, our literature review of *JAP* and *JCCP* identified two applications of an MNAR approach (Glisson et al., 2010; McKay et al., 2010). At the present time, these approaches are not yet implemented in common statistical packages. They have been shown to work well only under narrow ranges of conditions that are difficult to verify in practice. Consequently, MNAR approaches are unlikely to be widely applied in clinical psychology in the near future. MNAR approaches can serve a very useful purpose when used in conjunction with FIML or MI – if MNAR approaches produce similar results to those of MI or FIML, they can further increase our confidence that missing data are not materially affecting the results.

Nonindependent Data

Standard MR assumes that the observations are independent. This assumption is commonly violated when data are collected from groups (e.g., community groups, families) or when repeated observations are collected from the same set of individuals over time. When the observations are not independent, substantial inflation of the type I error rate can occur, making the results untrustworthy. Generations of researchers were taught to plan data collection to avoid dependency because of the problems in statistical analyses (e.g., selecting the data from only one child per family for use in analyses). Fortunately, extensions of MR have been developed over the past 25 years that not only correct for dependency in the data but also allow researchers to exploit the dependency and ask new questions of their data. Below we consider extensions of regression models for (a) group data (multilevel modeling) and (b) repeated measures collected over time (growth curve modeling). Although we consider the two cases separately, the statistical models used for their analysis are closely related (Mehta & West, 2000; Singer & Willett, 2003).

Group Data: Multilevel Modeling

Multilevel modeling (a.k.a. hierarchical linear modeling) is useful when data are hierarchically structured: multiple units at the lower level of a hierarchy (level 1) are collected into a unique single unit at the next higher level (level 2). As one illustration, in a randomized preventive trial (Jones, Brown, Hoglund, & Aber, 2010), individual elementary-school students (level 1) each attended one of several different schools (level 2). Schools were randomly assigned to receive a treatment or control program. As a second illustration, in a study of couples therapy by Anker, Owen, Duncan, and Sparks (2010), individual patients (husband, wife; level 1) were each part of a single married couple (level 2); each couple was seen by only 1 of 20 therapists (level 3). Although multilevel models can be applied to hierarchical data structures with more than two levels, we limit our consideration here to the commonly used two-level model.

Conceptually, multilevel modeling can be thought of as conducting a separate regression analysis in each group. Consider the following illustration of a school-level randomized trial simplified from Jones and colleagues' (2010) preventive trial mentioned above. A researcher recruits 100 students at each of 50 schools. The researcher measures each child's baseline level of aggression (Agg_{base}) at the beginning of the school year. During the year, each school is randomly chosen to receive either an Intervention program designed to reduce aggression ($T = +0.5$) or a Control program ($T = -0.5$)—the binary T variable representing the treatment condition is contrast-coded. Baseline aggression is centered around the mean of the entire sample, $Agg_{base\text{-}C} = Agg_{base} - \text{Mean}(Agg_{base})$. At the end of the

school year, each child's level of aggression ($Agg_{outcome}$) is measured. For each school j we could imagine estimating a separate regression equation for each of the students within each school:

$$Agg_{outcome-ij} = b_{0j} + b_{1j} Agg_{base-C-ij} + e_{ij} \text{ (level 1)} \quad (22)$$

In the level 1 equation we use the student-level covariate $Agg_{base\text{-}C}$ to predict $Agg_{outcome}$, each student's level of aggression at the end of the school year. The subscripts i corresponding to student within school and j corresponding to school are included to keep track of the case to which we refer. If we were to estimate Equation (22) separately for each of the 50 schools, we would have 50 intercepts b_{0j} and 50 slopes b_{1j} that would characterize the relationship between $Agg_{base\text{-}C\text{-}j}$ and $Agg_{outcome\text{-}j}$ in each separate school. This type of relationship is depicted by the gray lines in Figure 13.7A for 10 of the schools.

In our example, there is only one predictor at level 2, the treatment condition T. We can predict the regression coefficients b_{0j} and b_{1j} in each of the 50 schools based on T in a set of level 2 equations:

$$b_{0_j} = \gamma_{00} + \gamma_{01} T_j + u_{0_j} \quad (23a)$$

$$b_{1_j} = \gamma_{10} + \gamma_{11} T_j + u_{1_j} \quad (23b)$$

The term γ_{00} in Equation (23a) represents the mean intercept and γ_{10} in Equation (23b) represents the mean slope, respectively, across all of the schools. Given that Agg_{base} is centered, γ_{01} in Equation (23a) represents the difference between the Intervention and Control schools for the *average* child on baseline aggression. γ_{11} in Equation (23b) represents the mean difference in the slopes between Agg_{base} and $Agg_{outcome}$ in the Intervention minus Control schools. Each of the coefficients may be tested for statistical significance, with the tests of γ_{01} and γ_{11} being of most theoretical importance in this case. Figure 13.7B portrays the results. Given that the data have been centered, the test of γ_{01} represents the effect of treatment for the average child in the study on Agg_{base}. Once again, if Agg_{base} had *not* been centered, the test of γ_{01} would have represented the reduction in aggression of a child with a baseline level of aggression of 0, an effect that is unlikely to be of interest. Although not obvious from the multilevel regression equations (Equations 22, 23a, 23b), the test of γ_{11} depicted in Figure 13.7B represents a test of a baseline × treatment interaction.[9] This interaction is often an important effect in prevention trials in which the intervention is expected to have a larger effect (here, lower levels of aggression at posttest) as the child's level of risk increases. u_{0j} and u_{1j} are the residuals in the level 2 equations for the intercept and slope, respectively.

The multilevel model also provides estimates and significance tests of the variance of the residuals for the intercepts u_0 and the slope u_1. If these variances are significant, the possibility exists that other variables might be accounting for this variation at the school level. For example, schools located in low-income neighborhoods might have a higher level of aggression and a higher relationship between baseline and end-of-the-year aggression. Potential

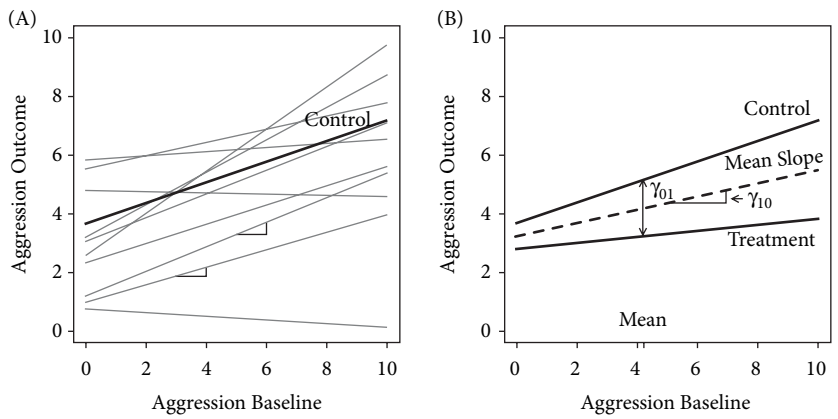


Figure 13.7 Multilevel modeling. (**A**) The light gray lines represent the slopes in 10 of the control schools. Only 10 schools are shown to avoid clutter. The heavy dark line represents the mean slope for all 25 control schools. (**B**) The dark lines represent the mean slopes for the 25 Control and 25 Intervention schools. The dotted line represents the mean slope (γ_{10}) for the Control and Intervention schools combined. γ_{01} represents the predicted mean difference between Intervention and Control programs at the mean of aggression in the sample at baseline.

level 2 explanatory IVs such as school neighborhood income could potentially be added to the level 2 regression equations.

In practice, the multilevel model is estimated in a single step using a procedure (maximum likelihood) that maximizes the efficiency of the estimates by using all of the available information. Programs that estimate multilevel models are available in standard statistical packages (e.g., SAS PROC MIXED, SPSS MIXED). A number of freestanding packages for multilevel analysis are also available (HLM, MLwiN). Hox (2010) and Snijders and Bosker (2011) provide accessible presentations of multilevel modeling.

Growth Curve Modeling

In growth curve modeling, repeated measures are taken on the same set of individuals over time. We present a simplified example below based on a study by Holmbeck and colleagues (2010). These researchers collected measures of adjustment (internalizing, externalizing behavior) on a sample of children every 2 years. At level 1, for each child there were four measurements taken at ages 8, 10, 12, and 14, denoted by the subscript *t*. At level 2, there were $n = 136$ children denoted by subscript *i*. The level 1 equation describes the linear trajectory for the development of each child's adjustment over the 6-year period.

$$ADJ_{it} = b_{0i} + b_{1i}Time_{it} + e_{it} \text{ (level 1)} \quad (24)$$

For ease of interpretation, $Time_{it}$ is normally set equal to 0 at the beginning of the study so the intercept can be interpreted as the child's initial status upon entering the study at age 8. Subsequent values of $Time_{it}$ represent the elapsed time since the beginning of the study, so that data were collected at values[10] of $Time_{it}$ = (0, 2, 4, and 6 years). ADJ_{ij} is the adjustment of each child *i* at $Time_t$. b_{0i} is the level of adjustment for child *i* at the beginning of the study (initial status) and b_{1i} represents child *i*'s linear rate of change per year over the 6-year period. Once again, we can imagine estimating b_0 and b_1 for each of the 136 children.

At level 2, the key individual difference variable was the child's developmental status (*DevStatus*). Half of the children ($n_{SB} = 68$) had a diagnosis of spina bifida and half ($n_{ND} = 68$) were normally developing children. Once again, we use a contrast coding scheme: *DevStatus* = –0.5 for the spina bifida group and +0.5 for the normally developing comparison group. We can write the level 2 equations predicting the initial status (intercept) and the slope of each of the children:

$$b_{0i} = \gamma_{00} + \gamma_{01}DevStatus_i + u_{oi} \quad (25a)$$

$$b_{1i} = \gamma_{10} + \gamma_{11}DevStatus_i + u_{1i} \quad (25b)$$

In Equation (25a), γ_{00} is the mean level of adjustment for all children at the beginning of the study since we used contrast coding. γ_{01} is the difference in the mean of the normally developing group minus the mean of the spina bifida group at the beginning of the study (initial status). In Equation (25b) γ_{10} is the unweighted mean of the slopes in the two *DevStatus* groups and γ_{11} is the difference between the mean slopes of the normally developing group minus the mean of the spina bifida group.

In practice, levels 1 and 2 are estimated simultaneously in a single step. These analyses can be performed using standard software for multilevel modeling (e.g., SAS PROC MIXED, SPSS MIXED) or structural equation modeling (e.g., Mplus). In some cases complications occur in growth curve modeling as the level 1 residuals may be correlated over time, and this feature may need to be included in the model. West, Ryu, Kwok, and Cham (2011) provide an introduction to latent growth curve modeling, and Singer and Willett (2003) provide a full-length treatment.

Generalized Estimating Equation

The generalized estimating equation approach (GEE) provides a second method of extending MR to address nonindependent data from groups or repeated measures. GEE focuses directly on correcting the error structure to represent the sources of nonindependence in the data. The results of estimates of the same regression problem using multilevel modeling and GEE are often quite similar. GEE makes weaker assumptions and is more robust to violations of its underlying assumptions than multilevel modeling; however, it makes the stronger and less realistic assumption that missing data are MCAR rather than MAR. The focus of GEEs is on estimating average effects in the population, whereas multilevel modeling also provides more information about individual participants and their variability. GEE is appropriate for designs like our example in which participants are measured at a fixed set of time points, whereas multilevel modeling can also address designs in which each participant is measured at a different set of time points. Ballinger (2004) provides an introduction and Hardin and Hilbe (2012) provide a full-length treatment of GEE.

Noncontinuous Dependent Variables: The Generalized Linear Model

MR and its special case ANOVA are statistical procedures that fall within a class of statistical models referred to as the *general linear model* (GLM). As we have discussed above, all applications of the GLM account for a dependent variable in terms of a set of IVs, whether the IVs are factors in an ANOVA, predictors in MR, or some combination of factors and predictors. All statistical procedures in the GLM framework require that the DV be continuous.

As discussed in an earlier section, MR makes assumptions about the characteristics of the residuals, referred to as the *error structure* of the analysis. It is assumed that (1) residuals exhibit homoscedasticity – that is, the variance of the residuals is constant for every value of the predicted value $\hat{Y}$; (2) residuals are independent of one another (addressed in the previous section); and (3) for accuracy of inference, residuals follow a specific probability distribution, the normal distribution. Also of importance, in the GLM the predicted scores are in the same units as the observed outcomes. For example, if the outcome is measured using the BDI, the predicted scores are in the units of the BDI. However, when the DV is not continuous, the units may differ so that a transformation known as a *link function* is needed to put the predicted and observed scores into the same metric.

Numerous dependent variables of interest are *not* continuous. To cite two common examples, DVs may be binary (e.g., Clinical Diagnosis: yes or no) or they may be counts, frequently with many zeros (e.g., symptom counts). If DVs such as these are analyzed with OLS regression, the residuals will *not* exhibit homoscedasticity, and they will *not* be normally distributed. The generalized linear model (GLiM) extends MR to address a wide variety of forms of noncontinuous DV by addressing assumptions (1) and (2) and by specifying the link function between observed and predicted scores. GLiM provides accurate estimates of regression coefficients, their standard errors, and tests of statistical significance. Several regression models, including logistic regression for binary and ordered categorical outcomes and Poisson regression for count DVs, are examples of varieties of regression models within the GLiM framework. As noted earlier in our review of studies in *JAP* and *JCCP*, they are commonly used by clinical researchers. Each of the regression models within the GLiM framework has a unique combination of an error structure and a link function (Coxe, West, & Aiken, in press).

Logistic Regression Analysis

Binary Outcomes. Logistic regression is most commonly used to analyze binary outcome variables (e.g., Case = 1, Non-Case = 0 for individuals who do vs. do not have a specific diagnosis). Logistic regression uses the binomial distribution as its error structure. The binomial represents the distribution of independent replications of a binary event (e.g., passing or failing a test) that has a fixed probability P. The variance of the binomial is *not* constant but rather resembles a football – it is largest when $P = 0.5$ and decreases as P moves toward 0 or 1. The predicted scores in logistic regression are not binary – we do not obtain predicted scores of 1 versus 0 for case versus non-case. Instead, the predicted score in logistic regression is the predicted probability $\hat{\pi}$ that an individual with a specific set of scores on the predictors will be diagnosed as a case. The predicted probability $\hat{\pi}$ can take on any value between 0 and 1. Given the discrepancy in the observed versus predicted scores, a link function is needed to transform the predicted into the observed scores. Equation (26) gives the logistic regression equation: the predicted probability is $\hat{\pi}$ a function of two IVs X_1 and X_2. The form of the equation is unfamiliar, since the predictor portion of the regression equation appears in an exponential function. As a consequence, scores on the predictors are not linearly related to the predicted probability:

$$\hat{\pi} = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}} \tag{26}$$

Now we apply the link function for logistic regression known as the logit to transform this equation to appear like a standard linear regression equation on the predictor side, resulting in Equation (27):

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = b_0 + b_1 X_1 + b_2 X_2 \tag{27}$$

Transforming the equation produces a new form of the predicted score, the *logit*. The odds are the ratio of the predicted probability of being a case to the predicted probability of not being a case, $\frac{\hat{\pi}}{1-\hat{\pi}}$. The logit is the natural logarithm of the odds, $\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$. The logit ranges from minus infinity to plus infinity as the probability ranges from 0 to 1. When the probability equals 0.5, the logit = 0. The use of the logit transformation produces the simple linear regression equation (Equation (27)) that

Table 13.4 Logistic Regression Analysis Predicting Whether a Doctor Judges that the Injured Athlete Can Versus Cannot Participate in the Next Season (*Participate*, Yes = 1, No = 0) as a Function of Weeks Between the Occurrence of the Injury and the Physician's Judgment ($Weeks_C$) and Whether the Athlete Has Completely Recovered from a Previous Injury (*Recovery*: Recovered, Not Recovered).

Predictor	*b*	*SE*	*Wald* χ^2	*Df*	*p*	*Exp*$^{(b)}$	*95% C.I. for Exp*$^{(b)}$	
							Lower	*Upper*
Intercept	0.871	0.297	8.605	1	.003	2.390		
$Weeks_C$	0.159	0.065	5.976	1	.015	1.173	1.032	1.332
Recovery	1.239	0.501	6.121	1	.013	3.451	1.294	9.208

mimics the right-hand side of a regression with a continuous DV; however, the logit is not a familiar metric to researchers. Fortunately, it is easy to transform from the logit back to the probability metric and describe the predicted probability $\hat{\pi}$ of being a case for each participant.

Equation (27), the linear form of logistic regression, has the same familiar regression coefficients and interpretation as OLS regression. The intercept b_0 is the value of the logit when all predictors equal zero. The regression coefficient b_1 is the increase in the logit for a 1-unit increase in the predictor X_1 holding X_2 constant; b_2 is the increase in the logit for a 1-unit increase in X_2 holding X_1 constant.

The relationship of the IVs to the DV is often expressed in a second way, particularly in journals in medicine and epidemiology. The *odds ratio* is the amount by which the odds are *multiplied* for a 1-unit increase in each IV holding the other IVs constant. Significance tests of the odds ratios and the corresponding linear regression coefficients b_0, b_1, and b_2 give identical test statistics and p values; they are transformations of one another. There is a simple relationship between odds ratios and familiar regression coefficients: if $b_1 = 0$, then the odds ratio = 1. If b_1 is positive, then the odds ratio is greater than 1. If b_1 is negative, then the odds ratio will be between 0 and 0.999.

We return to our earlier example of injured athletes. Suppose the DV were now the judgment by a physician at the beginning of the new season of whether the athlete can *Participate* on the team (Yes = 1; No = 0). *Participate* is predicted as a function of the number of weeks elapsed between the time of the injury and the physician judgment (*Weeks*) and *Recovery* (1 = Recovered; 0 = Not Recovered) from a previous serious injury. In all, 78 percent of the 120 athletes are judged ready to *Participate* (*Participate* = Yes, $n_P = 94$; *Participate* = No, $n_{NP} = 26$). There is a strong relationship between *Recovery* from the previous injury and physician's judgment: of those cleared to *Participate*, 56 percent of the 94 athletes have recovered from their previous injury; of those judged not able to *Participate*, only 27 percent of the 26 athletes have recovered from their previous injury. The correlation between the binary IV *Recovery* and the binary DV *Participate* is $\varphi = 0.243$, Pearson $\chi^2\,(1) = 7.07$, $p < .01$. Further, those judged ready to play have a longer elapsed time since the injury: they were assessed by the physician 12.0 weeks following the injury as opposed to 9.7 weeks for those not judged ready to play, $F(1, 118) = 7.30$, $p < .01$.

Table 13.4 presents the results of using Equation (27) to predict physician's judgment of *Participate* from centered $Weeks_C$ since the injury and *Recovery* from the previous injury.

$$\text{Logit}(\widehat{Participate}) = 0.871 + 0.159\,Weeks_C + 1.239\,Recovery \qquad (28)$$

The logit increases by 0.159 points for each week since the injury and by 1.239 points if the athlete has recovered from a previous injury. We note that rather than the familiar t tests of significance of these coefficients, Wald chi-square (χ^2) tests are reported, each with $df = 1$. The square root of each Wald χ^2 test is the familiar z test used in large-sample statistical tests. The Wald tests are asymptotic; they assume that we have the whole population. As shown in Table 13.4, both predictors are statistically significant.

The second form of the coefficients for the predictors, the odds ratios, is given in the column Exp(b). The odds ratio of 1.173 tells us that the odds that the doctor will judge that the athlete can participate are *multiplied* by 1.173 for *each week* that elapses between the injury and the date the doctor makes

the judgment. The conclusion is the same from the two sets of results. $b_1 = 0.159$ tells us that the logit of *Participate* increases with each passing week between the injury and the physician's judgment. The odds ratio tells us that the odds of *Participate* are multiplied positively (i.e., increase) with each passing week. The confidence intervals on the odds ratios are important and are commonly reported in journals. Recall that when the odds ratio equals 1, the regression coefficient b for the corresponding IV is 0, which indicates that the IV provides no unique contribution to the prediction of the DV. In Table 13.4, the confidence intervals for both $Weeks_C$ and *Recovery* do not include the value 1, also indicating a statistically significant effect. To illustrate, for *Recovery*, the odds ratio equals 3.451, and the confidence interval ranges from 1.294 to 9.208.

The test of overall significance of prediction from the logistic regression equation is no longer the familiar F test of R^2 used in OLS regression. It is a special likelihood ratio chi-square test (not the familiar Pearson chi-square) with degrees of freedom equal to the number of predictors, here 2. For our athletic participation example in Table 13.4, the likelihood ratio $\chi^2(2) = 13.936$, $p < .001$. Note that Equation (28) includes only $Weeks_C$ and *Recovery* as IVs. Paralleling our earlier analysis of player confidence presented in the section on interactions, we can also add the interaction term to the regression equation containing the two predictors:

$$\text{Logit}(\widehat{Participate}) = 0.820 + 0.103\,Weeks_C + 1.623\,Recovery + 0.238\,Weeks_C \times Recovery \quad (29)$$

The likelihood ratio chi-square test has increased to $\chi^2(3) = 16.277$, $p < .001$. However, we note that in this regression equation, the interaction is not significant, Wald $\chi^2(1) = 2.122$, $p = .145$. An alternative to the Wald chi-square test that often has greater statistical power at smaller sample sizes is the test of the difference in the likelihood ratio chi-squares comparing the full versus reduced models. This test follows the logic of the gain in prediction test used in MR. For the interaction term, this test is a likelihood ratio chi-square test of gain in prediction with degrees of freedom equal to the number of predictors added to the equation, here 1 – the interaction term. For our example, the likelihood $\chi^2(1) = 16.277 - 13.936 = 2.341$, $p = .126$. The failure to detect an interaction in the present example conveys an important message about the use of clinical diagnosis (1 = yes; 0 = no) as an outcome variable. Diagnosis is convenient, readily understood, and necessary when a clinical judgment is required (here, physician clearance for team participation). However, the use of a binary diagnosis instead of an underlying continuous rating (here, a physician rating) as the DV in the regression equation comes at a cost. Taylor, West, and Aiken (2006) demonstrate the substantial reduction in statistical power – the ability of detect a true effect if it in fact exists – when a binary diagnostic judgment replaces a continuous rating. Consistent with this general finding, although the interaction is not significant in the present example, the prediction of *Participate* from $Weeks_C$ is statistically significant in the group with full recovery from the prior injury ($b = 0.341$, odds ratio = 2.158, $p = .02$) but is not significant among those who have not had a full recovery ($b = 0.103$, odds ratio = 1.109, *ns*).

In logistic regression and other analyses included in the GLiM family, there is no simple R^2 measure of the standardized effect size for prediction from the whole regression equation. Because the variance of the residuals varies as a function of $\hat{\pi}$ in the binomial error distribution, the interpretation of proportion of variance accounted for becomes complex. Several analogues to R^2 have been proposed. One commonly reported measure is the Nagelkerke R^2, which ranges from 0 to 1 like familiar R^2 in MR. Cohen and colleagues (2003, Chapter 12, pp. 502–504) discuss this and other R^2 analogue measures. These measures have difficulties – for example, they do not necessarily increase monotonically as odds ratios increase, and many of them do not have a maximum of 1.

Multiple Categories: Unordered and Ordered. Logistic regression can be extended in two ways. One extension is to multinomial logistic regression, in which the DV consists of three or more unordered groups. Suppose in a longitudinal study a researcher identifies three groups of young adults: patients whose primary clinical symptoms are (a) severe anxiety, (b) severe depression, or (c) a comparison group of patients who do not have clinically significant symptoms of either problem (*Diagnosis*: 0 = Control, 1 = Anxiety, 2 = Depression). For each patient, whether the patient's biological mother or father had a history of a serious mental illness (*ParentHistory*) and a measure of the quality of the interaction with each patient with his or her parents during childhood (*Quality*) were available. A reference group (here, *Diagnosis* = Control) is defined for the DV. *Diagnosis* is partitioned into two contrasts: *Contrast A*, Anxiety vs. Control; *Contrast B*, Depression vs. Control. A separate logistic regression

equation is estimated for each DV contrast using the same set of IVs.

$$\text{Logit}(\textit{Contrast A}) = b_0(\text{A}) + b1(\text{A})\textit{ParentHistory} + b2(\text{A})\textit{Quality} \quad (30a)$$

$$\text{Logit}(\textit{Contrast B}) = b_0(\text{B}) + b1(\text{B})\textit{ParentHistory} + b2(\text{B})\textit{Quality} \quad (30b)$$

The two equations are estimated simultaneously and there is one overall likelihood ratio chi-square test for the full model. The values of each of the corresponding regression coefficients will typically differ across Equations (30a) and (30b). For example, the influence of *ParentHistory* controlling for *Quality* is unlikely to be the same for *Contrast A* (Anxiety vs. Control) and *Contrast B* (Depression vs. Control).

Consider now a case in which the outcome groups are ordered from lowest to highest. For example, a researcher might be interested in predicting which patients attend no sessions, some sessions, or all sessions of a clinical treatment program (Lester, Resick, Young-Xu, & Artz, 2010). Suppose the researcher uses *ParentHistory* and *Quality* as predictors. Ordinal logistic regression can be used in which it is assumed that the same relationship characterizes the transition from no to some sessions and from some to all sessions. In this case, Equations (31a) and (31b), which are similar to those used for multinomial logistic regression, would be used:

$$\text{Logit}(\textit{Contrast A}) = b_0(\text{A}) + b_1\textit{ParentHistory} + b_2\textit{Quality} \quad (31a)$$

$$\text{Logit}(\textit{Contrast B}) = b_0(\text{B}) + b_1\textit{ParentHistory} + b_2\textit{Quality} \quad (31b)$$

For ordinal logistic regression, Contrast A would compare the outcomes No with Some sessions and Contrast B would compare Some with All sessions. Unlike in Equations (30a) and (30b) for unordered multinomial outcomes, note that there is now a single b_1 and single b_2 that characterize both Equation (31a) and (31b). b_0(A) and b_0(B) still differ. b_0(A) represents the threshold on the predicted logit that must be crossed before a transition in categories from None to Some occurs and b_0(B) represents the threshold before the transition from Some to All sessions is made. As in binary logistic regression, the regression coefficients can be converted to odds ratios by exponentiation, odds ratio = e^b. Cohen and colleagues (2003, Chapter 13) provides a detailed introduction to logistic regression models and sample SAS and SPSS computer code.

Counts: Poisson Regression

Some clinical outcome variables involve a count of the behavior over a fixed period of time. Alcohol researchers might measure number of alcoholic drinks consumed per day; couples researchers might measure the number of incidents of intimate partner violence over a 6-month period. Count variables, particularly when the counts are low, violate the assumptions of standard OLS regression so that variants of the GLiM are needed. Two variants are normally considered, Poisson regression and negative binomial regression – the two variants typically yield similar hypothesis tests and each has its own strengths and limitations. Given space limitations, we only consider Poisson regression briefly here.

In count data, the variance of the residuals increases with the count. Poisson regression uses the Poisson distribution as the error structure for residuals. In a Poisson distribution the predicted mean and the variance are identical so that the variance increases with the predicted count. Second, counts are discrete (0, 1, 2, 3, . . .) and cannot take on negative values. As in logistic regression, a link function is needed to put the predicted scores into the same metric as the observed scores. In Poisson regression the link function is the natural logarithm of the predicted count of the DV, $\ln(\hat{\mu})$, where $\hat{\mu}$ is the predicted count. The linear form of the Poisson regression equation is:

$$\ln(\hat{\mu}) = b_0 + b_1X_1 + b_2X_2. \quad (32)$$

b_0 is the predicted value of the logarithm of the count when X_1 and X_2 both equal 0. b_1 is change in logarithm of the count for a 1-unit change in X_1, and b_2 is the change in the logarithm of the count for a 1-unit change in X_2, both holding the value of the other IV constant. Wald tests of the statistical significance of each regression coefficient and the likelihood ratio chi-square test of the full model may be performed.

This linear form of the Poisson regression equation is familiar and convenient statistically, but the predictions are in the unfamiliar metric of the natural logarithm. As with logistic regression, the Poisson regression equation may be written in a second exponential form that predicts the counts directly:

$$e^{\ln(\hat{\mu})} = e^{(b_0 + b_1X_1 + b_xX_2)} \quad (33)$$

In this metric, e^{b1} represents the amount by which the predicted count $\hat{\mu}$ is *multiplied* for a 1-unit change in X_1, holding X_2 constant. Coxe, West, and Aiken (2009) present a full description of Poisson regression and other alternatives within the GLiM for count data.

Summary and Conclusion

In this chapter we provided a broad overview of MR for clinical researchers. We considered the basics of multiple linear regression with both continuous and categorical IVs, and we noted that MR analysis requires that several assumptions be met. Although checks on these assumptions are not commonly reported in the published literature, they are important because they provide assurance that the regression coefficients, hypothesis tests, and confidence intervals will be correct. Articles in clinical psychology increasingly involve extensions of MR that confront the complexities of real-world clinical data that may violate the assumptions of the basic linear regression model. We considered the specification and interpretation of nonlinear and interactive relationships, methods of adjusting for missing data, methods for addressing nonindependent data, and methods for addressing DVs that are not continuous. Our overview introduced a variety of tools within the MR approach that permit clinical researchers to test interesting hypothesis with complex, real-world clinical data.

Notes

1. Path analysis with continuous outcomes was included in this total because it may be considered as a more general case of MR. In addition, analysis of variance (ANOVA) and analysis of covariance (25 studies, 54 percent) were included in this total since they can be seen as special cases of MR. Following the tradition of using ANOVA to analyze randomized experimental trials, studies in *JCCP* are more likely to report using ANOVA than MR.

2. Although latent trajectory growth modeling can be conducted in multilevel framework with identical results (Mehta & West, 2000), we separated this classification here from other applications of multilevel modeling.

3. Standardized solutions are not normally reported if any of the predictors are binary variables. The standardized solution is very sensitive to the proportion of cases in each group, whereas the unstandardized solution is not.

4. In regression equations containing an intercept, only G—1 dummy variables are needed. To illustrate, consider gender coded male =1, female = 0—here $G = 2$. If we know a person is *not* male, then the person must be female. A second dummy variable would provide information entirely redundant with the first.

5. Pairwise deletion programs typically make an adjustment in the degrees of freedom for purposes of significance testing.

6. The extent to which missing data affect the results of a study depends on both the proportion of missing data and the type of missing data (described shortly). A tiny proportion of missing data in a large sample is unlikely to materially affect the results; data that are missing completely at random do not bias the results, although they can still lower statistical power.

7. Statistical tests exist for data that are missing completely at random (see Enders, 2010). However, these tests make strong assumptions and are typically only of minimal value.

8. When FIML and MI are employed in the same dataset using the same set of variables, they typically produce very similar results. Theoretically, they produce hypothesis tests and confidence intervals that are asymptotically identical.

9. Equations (23a) and (23b) may be substituted into Equation (22) to produce a single reduced-form equation known as the mixed model equation. For this example, the mixed model equation is.

$$Agg_{outcome-ij} = \gamma_{00} + \gamma_{01} T_j + \gamma_{10} Agg_{base-C} + \gamma_{11} T_j x Agg_{base-C} + u_{0j} + u_{1j} Agg_{base-C} + e_{ij}$$

The interaction term $T_j Agg_{base-C}$ is more apparent.

10. $Time_{it}$ has two subscripts so that the model can consider cases in which each participant is measured at a different set of time points. For example, participant 1 might be measured at 0, 2, 4, and 6 years, whereas participant 2 is measured at 1, 5, and 8 years. In our example the data are collected at fixed measurement periods across all participants so the i subscript is unnecessary; $Time_t$ fully describes the measurement occasions.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Anker, M. G., Owen, J., Duncan, B. L., & Sparks, J. A. (2010). The alliance in couple therapy: Partner influence, early change, and alliance patterns in a naturalistic sample. *Journal of Consulting and Clinical Psychology*, *78*, 635–645.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*, 137–150.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality*, *59*, 143–178.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, *91*, 121–136.

Coxe, S., West, S. G., & West, L. S. (in press). Generalized linear models. In T. Little (Ed.), *Oxford handbook of quantitative methods*. New York: Oxford.

Doron-LaMarca, S., Vogt, D. W., King, D. W., King, L. A., & Saxe, G. N. (2010). Pretrauma problems, prior stressor exposure, and gender as predictors of change in posttraumatic stress symptoms among physically injured children and adolescents. *Journal of Consulting and Clinical Psychology*, *78*, 781–793.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.

Glisson, C., Schoenwald, S. K., Hemmelgarn, A., Green, P., Dukes, D., Armstrong, K. S., & Chapman, J. E. (2010). Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. *Journal of Consulting and Clinical Psychology*, *78*, 537–550.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Hardin, J. W., & Hilbe, J. M. (2012). *Generalized estimating equations* (2nd ed.). New York: Chapman & Hall/CRC.

Holmbeck, G. N., DeLucia, C., Essner, B., Kelly, L., Zebracki, K., Friedman, D., & Jandasek, B. (2010). Trajectories of psychosocial adjustment in adolescents with spina bifida: A 6-year, four-wave longitudinal follow-up. *Journal of Consulting and Clinical Psychology*, *78*, 511–525.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Jones, S. M., Brown, J. L., Hoglund, W. L. G., & Aber, J. L. (2010). A school-randomized clinical trial of an integrated social-emotional learning and literary intervention: Impacts after 1 school year. *Journal of Consulting and Clinical Psychology*, *78*, 829–842.

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*, 794–805.

Lester, K., Resick, P. A., Young-Xu, Y., & Artz, C. (2010). Impact of race on early treatment termination and outcomes in posttraumatic stress disorder treatment. *Journal of Consulting and Clinical Psychology*, *78*, 480–489.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

McKay, J. R., Van Horn, D. H., Oslin, D. W., Lynch, K. G., Ivey, M., Ward, K., Craplin, M. L., Becher, J. R., & Coviello, D. M. (2010). A randomized trial of extended telephone-based continuing care for alcohol dependence: Within-treatment substance use outcomes. *Journal of Consulting and Clinical Psychology*, *78*, 912–923.

Mehta, P., & West, S. G. (2000). Putting the individual back in individual growth curves. *Psychological Methods*, *5*, 23–43.

Nisenbaum, R., Links, P. S., Eynan, R., & Heisel, M. J. (2010). Variability and predictors of negative mood intensity in patients with borderline personality disorder and recurrent suicidal behavior: Multilevel analyses applied to experience sampling methodology. *Journal of Abnormal Psychology*, *119*, 433–439.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*, 1–25.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Schafer, J. L. (1999). *NORM: Multiple imputation of incomplete multivariate data under a normal model* [Computer software]. University Park, PA: Department of Statistics, Pennsylvania State University. Retrieved June 10, 2012, from http://sites.stat.psu.edu/~jls/misoftwa.html.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome is coarsely categorized. *Educational and Psychological Measurement*, *66*, 228–239.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, *64*, 1–48.

West, S. G., Ryu, E., Kwok, O.-M., & Cham, H. (2011). Multilevel modeling: Current and future applications in personality research. *Journal of Personality*, *79*, 1–50.

CHAPTER

14

Statistical Methods for Use in the Analysis of Randomized Clinical Trials Utilizing a Pretreatment, Posttreatment, Follow-up (PPF) Paradigm

Kendra L. Read, Philip C. Kendall, Mathew M. Carper, *and* Joseph R. Rausch

Abstract

Determining if a treatment "works" requires proper research design and statistical analysis. The randomized clinical trial (RCT) is the preferred research design to determine the efficacy of a given treatment. A variety of strategies exist for analyzing data from RCTs that follow a pretreatment, posttreatment, follow-up (PPF) design. This chapter reviews common data analytic approaches and discusses the relative advantages and disadvantages of each. The chapter also reviews when to apply each of these strategies for analyzing data within a PPF design. Analyses reviewed include the analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), analysis of covariance (ANCOVA), multivariate analysis of covariance (MANCOVA), and hierarchical linear modeling (HLM).

Key Words: Randomized clinical trial (RCT), clinical research, analysis of variance (ANOVA), analysis of covariance (ANCOVA), multivariate analysis of covariance (MANCOVA), hierarchical linear modeling (HLM)

The randomized clinical trial (RCT) has been established as a premier tool with which to evaluate whether a treatment qualifies as empirically supported. The RCT methodology can apply to preventive interventions seeking to reduce risk for disorder, interventions that strive to enhance one's life, or therapeutic interventions, where treatments are evaluated for persons with identified/diagnosed conditions. When conducting an RCT, there are a number of analytic options available with which to evaluate your results. Given the variety of statistical options, how do you choose the best one for your study? Decisions about statistical methods are intricately linked to the study design and to the research questions being asked. Clarity about methodology (e.g., measurement and timing, randomization, structure of the conditions) will guide appropriate choices of statistical tests (see Chapter 4 in this volume). When more than one statistical option exists, one must consider the underlying assumptions made by each class of statistical test, as well as the resultant variations in power and precision they bring to one's conclusions. One strives to choose the optimal test to minimize the chance of biased parameter estimates and to maximize power and precision in the evaluation of differences.

This chapter provides an overview of statistical methods that may be applied to continuous data in the context of pretreatment, posttreatment, follow-up (PPF) designs, which are typically used for RCTs evaluating interventions. Within this research paradigm, a particular outcome variable is measured in each treatment condition (e.g., Treatment A, Treatment B, Control Condition) at three different time points: before inception into treatment (pre), immediately after treatment has concluded (post), and at a specified and standardized point after the conclusion of treatment (follow-up; e.g., 1-year follow-up assessment). This particular structure is often used in clinical trials to assess changes across conditions over time, including immediate effects (seen at posttreatment) and the lasting benefits of

the particular intervention (seen at follow-up). This chapter reviews and compares common analytic strategies for RCT data that seek to answer typical questions about change across conditions over time in intervention research.

General Assumptions

In general, analytic strategies for the PPF RCT extend from the more basic pre–post design, in which a dependent variable is measured at just two time points. With one substantial exception (examination of differences from post to follow-up, discussed later), the majority of the basic assumptions and analytic strategies will be the same when analyses extend to inclusion of the follow-up time point. For example, the same recommendations hold for analyses examining differences between conditions from pre to post as from pre to follow-up. Multivariate analyses that consider both post and follow-up time points as concurrent dependent variables will also be discussed.

For the present purposes, a number of general assumptions must be met for all statistical analyses considered. First, experimental conditions must be the result of random assignment (as required in experimental design). The purpose of random assignment to conditions is to increase the likelihood of baseline equality in the between-condition variable (outcome variable), affording causal inferences about differences in those conditions at posttreatment or follow-up. For example, in a study of the effects of two treatments and a control condition on anxiety severity, random assignment increases the chances that the three conditions are equal/comparable prior to initiation of treatment in terms of the mean anxiety severity, age, and other important variables. In addition to random assignment, researchers may match participants between conditions in accordance with their pretreatment scores and other covariates (Friedman, Furber, & DeMes, 1998; Matthews, 2000). Although variations of these analytic methods may be applied to nonrandomized designs, this chapter will use random assignment (as in an RCT) as an assumption for all statistical plans and considerations.

There are a number of additional assumptions necessary when considering the use of parametric tests:

- The dependent variable should be normally distributed within each treatment condition, and for analysis of covariance only, conditional on the covariate (most frequently this will be the pretreatment scores).
- Statistical independence of participant scores at each time point. Thus, scores on dependent variables represent only one participant at that time point.
- Homogeneity of variance (i.e., the variance of the dependent variable is equal/comparable for all conditions, such that one condition does not have more variable scores than another). Heterogeneity of variance could provide biased estimates of parameter means and reduce the power of a given statistical test.

It is also important to consider the influence of missing data for the specific type of statistical test you plan to run. RCTs can have trouble with missing data: we cannot ethically make every participant stay with the study to complete the treatment course to which he or she was assigned. It is important to examine the aspects of study design and statistical analysis that can reduce the unwanted effects of missing data. One promising point is that potential bias from missing data is reduced when data are missing at random, instead of systematically. This statement assumes that conditions do not differ in the degree and nature of their missing data (e.g., one particular condition is not missing substantially more data than others or missing only individuals of a particular socioeconomic background or severity rating). Patterns of missingness should be evaluated to identify nonrandomly missing data that could pose problems for inferential statistics. In some cases, it is possible to adjust condition means for comparison based on differences at pretreatment (analysis of covariance), which results in less bias from missing data. There are generally no hard-and-fast guidelines for handling missing data within particular datasets, but generally less than 5 percent missing data for any particular cell produces minimal damage to future analyses. For more advanced consideration of the appropriate handling of missing data, the reader is referred to Chapter 19 in this volume.

In general, the purpose of these guides is to help you choose a single statistical method that minimizes biased estimates and maximizes power and precision for the particular research question and design. Comparisons are made between particular analytic methods throughout this chapter, but it is not recommended that researchers perform all strategies discussed in a horserace-type fashion, reporting the test that offers the most favorable result. Multiple statistical runs of this sort inflate the chance of type I error (i.e., the chance of falsely rejecting

the null hypothesis or "finding" a difference when no true difference exists in the population). Thus, choice of statistical methods, and the reporting of their respective statistics and significance values, is best made *a priori*.

Analysis of Variance

Tests of analysis of variance (ANOVA) allow us to examine mean differences between several groups, especially when several independent variables are included in the model. Within a PPF RCT design, ANOVA tests can be used to test the null hypothesis that the means of all conditions are equal across time. With ANOVA, we can explore omnibus main effects of potential differences between conditions and differences across time, as well as potential interactions between these conditions and time. Each potential test is examined in turn below.

The main effect of time considers differences between pre and post or pre and follow-up, averaging across conditions. This test examines the null hypothesis that mean scores of each time point, averaged across treatment conditions, are equal. This test allows us to consider whether there was an overall decrease or increase in scores over time, but does not tell us anything about differences between conditions. As such, alone, it is not likely to be very useful for a treatment outcome study, in which the impetus is to explore differences in a dependent variable between treatment conditions. For less rigorous trials not including a comparison group, evaluation of the main effect of time constitutes the only available interpretative tool.

The main effect of condition considers differences between treatment conditions, averaging across time. It tests the null hypothesis that the condition means, averaged across time points, are equal. The mathematical model for the test of main effect of treatment condition can be expressed by equations 1 (pre to post) and 2 (pre to follow-up) in the appendix.

Omnibus ANOVA tests have the advantage over some other statistical approaches because they test for overall differences between three or more conditions, but they are often not the most powerful test to consider for PPF designs. As is shown in ANOVA equations 1 and 2, the omnibus test of the main effect of treatment condition restricts the regression slope from pre to post/follow-up to be −1 (see Rausch, Maxwell, & Kelley, 2003, for mathematical derivation). Such a rigid mathematical assumption decreases the power of ANOVA main effects tests to find differences between conditions for which the regression slope is positive or of any other magnitude than 1. In sum, the rigid regression slope assumptions of ANOVA tests of main effects are unlikely to be met in practice.

ANOVA may also be used to evaluate a potential interaction between condition and time. This procedure may be used for research questions that want to examine whether conditions change differently between two time points, or whether condition means on the dependent variable are different at posttreatment or follow-up (in comparison to pretreatment). Stated differently, analyzing the interactive effect of condition and time allows the researcher to evaluate whether the effect of time is consistent across conditions, and conversely, whether the effect of condition is consistent across time points. Specifically, interaction tests evaluate the null hypothesis that condition means are equal at post/follow-up or that mean condition differences between pre and post/follow-up are equal. With the assumptions that assignments to treatment condition were randomized (and participants are thus theoretically equal on pretreatment measures) and that collection of data representing the pretreatment condition actually occurred prior to the start of treatment, these two questions are theoretically equivalent. The difference comes in the construction of the dependent variable: you can examine differences between mean scores at post/follow-up or difference scores (mean trajectory) between pre and post/follow-up. Test conclusions may be the same if conditions are equal on pretreatment measures (individual differences are controlled) and their slopes of change are relatively parallel (equal condition mean difference scores). However, tests of the difference scores that account for pretest differences may be more powerful when conditions differ on pretreatment measures. The mathematical model for the ANOVA time-by-condition interaction test can be expressed by equations 3 (pre to post) and 4 (pre to follow-up) in the appendix.

Unlike the test of main effects of treatment condition, the parameters of these ANOVA tests restrict the slope of the regression equation to 1. Although it may be more reasonable of an assumption to have a positive regression slope (i.e., a positive correlation between pre and post scores), there is likely no theoretical reason to restrict the magnitude of the slope to 1. Again, we must conclude that impractical regression slope restrictions lead to decreases in power and precision of these ANOVA tests when estimating population parameters. Although an ANOVA on the difference score answers a slightly

different question than the time-by-condition interaction, the mathematical results will be equivalent from a randomized PPF design (Huck & McLean, 1975), and the same shortcomings may be applied (Rausch, Maxwell, & Kelley, 2003).

Multivariate Analysis of Variance

One of the benefits of PPF designs is that one can examine change across multiple time points. An extension of ANOVA for multiple dependent variables or time points (MANOVA) allows for researchers to test omnibus condition-by-time interaction with multiple dependent time points in the same mathematical vein as the ANOVA tests discussed above. As with any statistical test, it is important to include only dependent variables for which there is clear theoretical basis, despite the increased abilities of this test. Additional assumptions must be met for this multivariate test, including:

- Multivariate normality: normal distribution of collective dependent scores within each treatment condition
- Relative homogeneity of covariance matrices: variances of each condition are roughly equal for each dependent variable, as are the correlations between each pair of dependent variables

Given that MANOVA is an extension of ANOVA for multiple dependent variables, the same restrictions and caveats apply as with ANOVA. Accordingly, it is recommended that researchers continue with adjustments of ANCOVA (MANCOVA) for multiple time points (discussed in later sections).

Analysis of Covariance

Discussion of the use of analysis of covariance (ANCOVA) is important given that, historically, it has often been commonly overlooked or maligned due to research misuse. ANCOVA was first described by Fisher (1932) as a means of more precisely analyzing data from randomized experiments to identify treatment effects (Maxwell, O'Callaghan, & Delaney, 1993). From its inception until the 1960s, researchers used it as a method to test for condition differences in nonrandomized studies, although this technique is no longer recommended (Campbell & Boruch, 1975; Cronbach & Furby, 1970; Elashoff, 1969). It has been proposed that ANCOVA fell out of favor following skepticism about adjusted means and the probability of meeting strict requirements needed for this procedure to alleviate the bias of nonrandomized comparisons (Cronbach, Rogosa, Floden, & Price, 1977; Maxwell, O'Callaghan, & Delaney, 1993; Reichart, 1979). However, historically negative attitudes about the use of ANCOVA overlook the benefits of using this particular analytic strategy within true randomized treatment experiments. Generally, ANCOVA represents a more powerful and precise statistical test than ANOVA when used for randomized treatment studies.

ANCOVA has been heralded as the most appropriate analytic method for examining condition differences across time in most RCTs (Huck & MacLean, 1975; Rausch, Maxwell, & Kelley, 2003). While still testing for differences between conditions or differences in change scores on a dependent variable, one unique aspect of ANCOVA is that it controls for differences between conditions on the initial assessment at pretreatment. Additionally, it allows for the data to estimate the relationships between the dependent variable and the covariate, typically increasing statistical power over an analysis that constrains this relationship to be a particular value (i.e., ANOVA). For PPF designs, ANCOVA is able to answer both kinds of research questions (differences at post/follow-up or differences in change between two time points [including pre]) when studies meet assumptions of random assignment to treatment condition and independence of observation between variables and covariates and dependent variables.

The primary benefit of ANCOVA is that it allows researchers to account for differences in at least one covariate, which is an additional variable that is significantly correlated with the dependent variable. Within the bounds of a PPF design, the covariate represents the measure of the dependent variable (e.g., severity of disorder) at pretreatment. It is important that covariates are measured before treatment begins; otherwise, differences between conditions on the first assessment—which could represent potential effects of the treatment—will be adjusted or equaled between conditions. Such a circumstance would greatly reduce the power and interpretation of test results; adjustment based on treatment-influenced covariates would result in removal of some of the treatment effect between conditions (Tabachnick & Fidell, 2007). In cases in which the assumptions of randomization and pretreatment measurement of the covariate have been met, ANCOVA analyses are relatively robust to the influence of measurement bias and nonlinear treatment effects, which otherwise might decrease the power of the test.

Thus, ANCOVA allows the researcher to test for condition differences on a dependent variable while adjusting for the linear effects—and subsequent

differences between conditions—in pretest scores (covariate). The inclusion of covariate adjustment with ANCOVA reduces differences between conditions at the start, increasing the ability to explore the true treatment differences without individual noise or "unhappy randomization" (Kenny, 1979, p. 217; Rausch, Maxwell, & Kelley, 2003), in which conditions happen to be unequal on the dependent variable pretest, despite the best efforts of randomization. Importantly, covariate adjustments reduce within-condition error (Field, 2005), allowing researchers to confidently state that any significant tests result from differences due to the effects of the treatment (Rausch, Maxwell, & Kelley, 2003).

In addition to the statistical assumptions mentioned in the previous sections, ANCOVA analyses require an additional assumption of homogeneity of regression slopes. Relative equivalence of regression slopes between the dependent variable and covariate is needed across conditions to make accurate comparisons. However, it is possible that theory would predict unequal regression slopes for different treatment conditions. In such cases, it is recommended that those specific parameters are included in the model (see Rogosa, 1980). The power of ANCOVA is likely greatest when there exists a linear relationship between variables included in the model (Rausch, Maxwell, & Kelley, 2003). It is possible to adjust for quadratic or cubic relationships with ANCOVA, but these techniques are not typically used for RCTs. When statistical assumptions of ANCOVA are met, this procedure is able to control for both the unconditional type I error rate (chance of falsely rejecting the null hypothesis when conditions are equal in the population after repeated tests), as well as the conditional type I error rate (chance of falsely rejecting the null hypothesis when conditions are truly equal after repeated tests, conditional on the same adjusted covariate values; Maxwell, 1994; Maxwell & Delaney, 1990; Senn, 1989). The mathematical model for ANCOVA for examining omnibus differences between treatment conditions can be found in equations 5 (pre to post) and 6 (pre to follow-up) in the appendix.

Notably, rather than fixing the slope predicting posttreatment (or follow-up) scores from pretreatment scores, ANCOVA *estimates* the population regression slope by using the correlation between pretreatment and posttreatment scores and the standard deviation of scores at each time point. This difference from the restrictive regression slope of ANOVA (either 1 or –1) represents an important advantage. ANCOVA's estimated regression slope results in substantial reductions in model error variance given that it allows the slope to reflect true patterns in the data. This reduction in error variance allows for important increases in power and precision that make ANCOVA the preferred statistical method for analyzing differences between conditions at posttreatment or follow-up time points (both with pretreatment as the covariate) in randomized designs. It is important to note that in addition to analyses of differences between conditions at either posttreatment or follow-up, one can also conduct an ANCOVA on the difference score between either time point and pretreatment (e.g., posttreatment–pretreatment; follow-up–pretreatment). The statistical results and inferential conclusions of these tests will be identical, as long as pretreatment is used as the covariate in both approaches2 (posttreatment–follow-up comparisons will be discussed in later sections). Choice between these methodologies is then left to the interpretive preferences of the researcher (Hendrix, Carter, & Hintze, 1979). ANCOVA analyses that include pretreatment as a covariate will be more powerful than respective ANOVA, even when the pretreatment is included as a linear component of the model.

For RCTs, covariates will likely represent preintervention assessments of dependent variables (e.g., symptom severity). However, there are a few important theoretical and statistical considerations when identifying covariates. First, covariates should be independent of treatment (i.e., gathered before initiation of treatment) and should be a guided by theory, just like one's choice in dependent variable. For example, one would not select head size as a covariate (or dependent variable) in studies aiming to examine treatment effects of a depression study. Covariates, like other variables, should be reliably measured or risk a decrease in power and an increase in chances of type II error (chance of falsely accepting the null hypothesis or finding no difference when a true differences exists between conditions in the population). Further, it is important that potential covariates are correlated with the dependent variable (e.g., symptom severity at posttreatment or follow-up) but are not highly correlated with each other. The inclusion of each covariate results in the loss of one degree of freedom of error, such that multiple covariates inflate the model error term. When multiple permutations are under consideration, covariates may be statistically evaluated against one another in repeated ANCOVAs to maximize efficiency. However, it is recommended that covariates be chosen *a priori*, based on theory and

prior studies. It is important to make parsimonious decisions about the incremental utility of additional covariates (represented in maximum adjustment of the dependent variable) to maximize the power and precision of one's model.

Multivariate Analysis of Covariance

Similar to multivariate adjustments of ANOVA (i.e., MANOVA), multivariate analysis of covariance (MANCOVA) allows for step-down tests of significant contributions of multiple dependent variables. In these comparisons, competing dependent variables are treated as covariates for one another. There are generally two different approaches for extending ANCOVA for multiple dependent variables. The first MANCOVA method examines posttreatment and follow-up time points in the model at the same time, with pretreatment as the covariate. The second technique allows for two simultaneous RCT comparisons to be made, with pretreatment as the covariate: condition differences in (1) the mean of posttreatment and follow-up scores on a dependent variable [*M*] and (2) the difference between posttreatment and follow-up scores [*D*]. Ultimately, the results of these two methods will provide the same omnibus results, as long as the pretreatment score is designated as the covariate in each model.

When significant omnibus results are returned from these analyses, planned pairwise contrasts will be important to identify specific differences between multiple conditions. Following MANCOVA omnibus tests, four types of ANCOVA pairwise comparisons may be used to identify specific condition differences, using pretreatment as a covariate: condition differences at posttreatment, condition differences at follow-up, mean of posttreatment or follow-up scores (*M*), and the difference between posttreatment and follow-up scores (*D*). ANCOVA analyses are generally recommended as clearer and more powerful follow-up comparisons for randomized PPF RCT designs, and this difference is reflected in smaller and more precise confidence intervals for pairwise comparisons (see appendices of Rausch, Maxwell, & Kelley, 2003, for mathematical comparisons). Again, it is **not** recommended that researchers try every technique in order to uncover favorable results; rather, pairwise comparisons should be logical and guided by theory. It may not be necessary to use a multiple comparison adjustment when using different varieties of the following pairwise comparisons, whereas some might argue that these questions can be considered distinct classes of questions that do not compromise the (type I) family-wise statistical error rate. However, it is important to use a multiple comparison adjustment (e.g., Bryant-Paulson, 1976, for ANCOVA) when calculating multiple pairwise comparisons of the same kind.

Researchers may also be interested in examining whether conditions differ on the average of their posttreatment and follow-up scores. Tests of averages can be more powerful than tests of individual time points. However, the use of this particular strategy depends on whether the slopes are relatively parallel between conditions, such that the average does not remove important change information. Thus, when the change or *D* variable is too similar between groups, it may be more useful to examine the condition differences on *M*.

Comparisons between Posttreatment and Follow-up. Thus far, we have examined differences between pretreatment and other time points in the PPF design. Researchers may seek to compare differences between posttreatment and follow-up to examine how treatment gains are maintained relative to other conditions. Interestingly, we expect conditions to be different at posttreatment if active treatments were implemented as part of the research design. Thus, we are comparing conditions we assume to be unequal at the start. There are two methods for handling this paradox (Lord, 1967). The different methods used may result in different or conflicting results, so it is important to consider the specific recommendations for each method.

The first method involves altering the equation for *D* that includes pretreatment as the covariate in the model (see equation 7 in the appendix). Results from use of this model must be interpreted with the caveat of potential differences at posttreatment on the covariates and/or dependent variable. Potential differences between conditions at post are considered within the *D* variable, which restricts the regression slope that predicts follow-up from post to a value of 1. This model allows researchers to examine whether conditions change differently between posttreatment and follow-up, or whether the magnitude of treatment effect is the same at posttreatment and follow-up time points. It cannot elucidate whether condition differences would be identified if conditions were equal at posttreatment. Following significant omnibus *D* ANCOVA condition comparison between posttreatment and follow-up, pairwise comparisons can be elucidated by examining the overlap of confidence intervals of change for each condition with each other and zero.

The second method includes both pretreatment and posttreatment as covariates in the model and examines whether the mean condition change would differ if adjusted for equality at posttreatment (see equation 8 in the appendix). The advantage of this model is that it allows for an estimated regression slope to predict follow-up data from post, rather than restricting it to 1. Again, we are unable to say whether treatment effects would remain if conditions were equal on the covariates and dependent variable at posttreatment. Generally, this model is particularly susceptible to the influence of measurement error, and is less useful in psychology studies where some degree of measurement error is expected.

Hierarchical Linear Modeling

Hierarchical linear modeling (HLM), also known as multilevel modeling and mixed-effects modeling (Raudenbush & Bryk, 2002), represents a somewhat more advanced statistical technique that allows for exploration of randomized longitudinal research studies. Although it is another technique for looking at randomized PPF RCT data, it has some important distinctions from the general linear model techniques of ANOVA and ANCOVA. Specifically, HLM allows for the examination of individual growth curves, and thus condition change, over time. This procedure simultaneously examines within-subject variations in repeated measurements of the dependent variable over time (level 1), as well as between-subject interactions of condition membership and the dependent variable over time (level 2; Gibb, Beevers, Andover, & Holleran, 2006). It is the combination of these two levels of analysis that makes up a mixed-effects model. Repeated assessment provides a distinct advantage over analyses of averaged assessments in ANOVA and ANCOVA comparisons (Cohen & Cohen, 1983). Averaged assessments can underestimate the variance of the model and overestimate the precision of statistical tests (DeRubeis et al., 2005; Gibb et al., 2006; Kreft & De Leeuw, 1998). More traditional methods described previously do not allow for this type of data analysis, leading to increased power and precision for finding condition differences for some research designs when using an HLM approach.

Linear growth curve modeling using PPF designs may be problematic if the true relationship in the population is not linear. If the true relationship is linear, then HLM is more powerful than ANOVA and ANCOVA, assuming that measurement time points are unequally spaced and that HLM uses the pretreatment as a covariate in level 2 (slope as the dependent variable; Rausch & Maxwell, 2003). Theoretically, we may not expect regression slopes of treatment response to be linear, especially in the absence of active treatment between posttreatment and follow-up time points. However, while a linear growth model may not reflect the true population relationship, it is relatively robust for short intervals of time and few observations (like the three included in PPF designs). Within HLM, it is possible to design polynomial functions of time (i.e., quadratic, cubic, quartic) to better model growth curves that reflect more accurate trajectories. However, this is rarely done within PPF designs, given that only three time points are usually available.

Summary

There are several methods for analyzing data from randomized PPF treatment outcome data from an RCT. Researchers are encouraged to examine differences in research designs and statistical techniques to maximize their power for finding true differences between treatment conditions. General limitations of all statistical analyses apply. Specifically, results from any study are not able to generalize beyond the populations that the samples are purported to represent within a given study. Researchers are encouraged to thoroughly examine their research designs, including sample size and composition, to ensure that the results of their analyses are appropriately applied to general populations.

Notes

1. A comprehensive review of the strategies to handle missing data is beyond the scope of this chapter. Readers are referred to Allison (2009), Tabachnick & Fidell (2007), Schafer & Graham (2002), and Sinharay, Stern, & Russell (2001) for more information.

2. Mathematical derivations and comparisons of these statistical methodologies, including calculations and comparisons of power, are beyond the scope of this chapter. For more information, see Rausch, Maxwell, & Kelley, 2003, Appendices A and B.

References

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.) *The Sage handbook of quantitative methods in psychology*. (pp. 72–89). Thousand Oaks, CA: Sage Publications Ltd.

Bryant, J. L., & Paulson, A. S. (1976). An extension of Tukey's method of multiple comparisons to experimental designs with concomitant variables. *Biometrika*, *63*, 631–638.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In

C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, *74*, 68–80.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. C. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias.* Paper presented at the Stanford Evaluation Consortium, Stanford, CA.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive therapy vs. medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, *62*, 409–416.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal*, *6*, 383–401.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Edinburgh: Oliver and Boyd.

Friedman, L. M., Furberg, C., & DeMets, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer-Verlag.

Gibb, B. E., Beevers, C. G., Andover, M. S., & Holleran, K. (2006). The hopelessness theory of depression: A prospective multi-wave test of the vulnerability-stress hypothesis. *Cognitive Therapy and Research*, *30*, 763–772.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1979). A comparison of five statistical methods for analyzing pretest–posttest designs. *Journal of Experimental Education*, *47*, 96–102.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest–posttest design: A potentially confusing task. *Psychological Bulletin*, *82*, 511–518.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305.

Matthews, J. N. S. (2000). *An introduction to randomized controlled clinical trials*. London: Arnold.

Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest-posttest designs. *Psychological Bulletin*, *115*, 142–152.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

Maxwell, S. E., O'Callaghan, M. F., & Delaney, H. D. (1993). Analysis of covariance. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 63–104). New York: Marcel Dekker.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rausch, J. R., & Maxwell, S. E. (2003). Longitudinal designs in randomized group comparisons: Optimizing power when the latent individual growth trajectories follow straight-lines. Manuscript in preparation.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology*, *32*, 467–486.

Reichart, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Boston: Houghton Mifflin.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Senn, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, *8*, 467–475.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317–329.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education, Inc.

Chapter 14 Appendix

Equations for ANOVA Main Effects

Equation 1: ANOVA Pre to Post Condition Main Effect:

$$Post_{ij} = \mu_{Post_i} + (-1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Equation 2: ANOVA Pre to Follow-up Condition Main Effect:

$$F/U_{ij} = \mu_{F/U_i} + (-1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Where Pre_{ij} is the pre score for the individual i in condition j; μ_{Pre} is the population grand mean of the dependent variable (e.g., anxiety severity) at pre; μ_{Postj} and $\mu_{F/Uj}$ are the population grand means of the dependent variable at post and follow-up, respectively, for condition j ($j = 1, 2, \ldots, a$, where a is the total number of conditions); and ε_{ij} is the error for the individual i ($i = 1, 2, \ldots, nj$, where n_j is the sample size in condition j) in condition j.

Equations for ANOVA Interactions

Equation 3: ANOVA Pre to Post Time by Condition Interaction Effect:

$$Post_{ij} = \mu_{Post_i} + (1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Equation 4: ANOVA Pre to Follow-up Time by Condition Interaction Effect:

$$F/U_{ij} = \mu_{F/U_i} + (1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Equations for ANCOVA Pre to Post/Follow-up Analyses

Equation 5: ANCOVA Pre to Post:

$$Post_{ij} = \mu_{Post_i} + \beta_{Post,Pre}(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Equation 6: ANCOVA Pre to Follow-up:

$$F/U_{ij} = \mu_{F/U_i} + \beta_{F/U,Pre}(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Equations for ANCOVA Post to Follow-Up Analyses

Equation 7: ANCOVA Post to Follow-up (Pre as covariate)

$$D_{ij} = \mu_{D_i} + b\ \beta_{D,Pre}(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij}$$

Where μ_{Dj} is the population mean of D for condition j;
$\beta_{D,Pre}$ is the population regression slope predicting D from the pretest; and
ε_{ij} is the error for individual i in condition j (Raush, Maxwell, & Kelley, 2003).

Equation 8: ANCOVA Post to Follow-up (Pre and Post as covariates)

$$F/U_{ij} = \mu_{F/U_i} + \beta_{F/U,Pre}(Pre_{ij} - \mu_{Pre}) + \beta_{F/U,Post}(Post_{ij} - \mu_{Post}) + \varepsilon_{ij}$$

Where μ_{Post} is the population grand mean at post;
$\mu_{F/Uj}$ is the population mean score on the follow-up for condition j; and
$\beta_{F/U,\ Pre}$ and $\beta_{F/U,\ Post}$ are the population partial, unrestricted regression slopes for pre and post, respectively (Rausch, Maxwell, & Kelley, 2003).

CHAPTER

15 Evaluating Treatment Mediators and Moderators

David P. MacKinnon, Ginger Lockhart, Amanda N. Baraldi, *and* Lois A. Gelfand

Abstract

This chapter outlines methods for identifying the mediating mechanisms by which treatments achieve effects and moderators of these effects. Mediating variables are variables that transmit a treatment effect to an outcome variable. Moderating variables are variables that identify the subgroups, conditions, or factors for which the treatment effect on the outcome differs. Reasons for conducting mediation and moderation analyses in treatment research are provided along with simple examples of mediation and moderation models. More detailed mediation models are described that incorporate multiple mediators, longitudinal measurement, experimental designs, and alternative approaches to causal inference for mediation. A research design for an exemplar treatment study that includes investigation of mediating and moderating processes is described. However, we acknowledge that the search for mediating and moderating variables in treatment research must be a multifaceted approach that includes information from a variety of sources in addition to ideal experimental design and careful measurement of constructs. The chapter concludes with future directions in mediation and moderation methodology in treatment research.

Key Words: mediation, moderation, indirect effect, mechanisms of treatment

> *"Rapid progress in identifying the most effective treatments and understanding on whom treatments work and do not work and why treatments work or do not work depends on efforts to identify moderators and mediators of treatment outcome. We recommend that randomized clinical trials routinely include and report such analysis."* (Kraemer, Wilson, Fairburn, & Agras, 2002, p. 877)

Early clinical research investigated whether treatments even worked (Paul, 1967). A historically sole focus on case study and studies without a comparison group has been replaced by a more rigorous and informative focus on studies comparing the change between a group receiving treatment and a group that did not receive treatment (or received an alternative treatment) (see Chapter 4 in this volume). With the discovery of successful treatments for a variety of disorders in these experimental designs, clinical science has matured to look at two common aspects of treatment: "How does the treatment achieve change?" and "Does the treatment lead to different changes for different people or in different contexts?" These two questions focus on mediating variables (the process by which treatment effects are achieved) and moderating variables (variables for which a treatment effect differs at different values of the moderating variable). Mediating variables are relevant in understanding how treatments work; moderating variables are relevant in understanding if treatment effects differ across individuals' characteristics or treatment contexts. With attention to mediation and moderation effects, researchers can enrich theoretical perspectives, extract more information from a research study, and provide stronger, more tailored treatments.

A growing number of treatment researchers have called for more attention to mediation to test the theory underlying treatments and to identify treatment actions that are effective (Kazdin, 2000; Kazdin & Nock, 2003; Longabaugh & Magill, 2011; Nock, 2007; Weersing & Weisz, 2002). As noted by Weisz and Kazdin (2003, p. 445), "The job of making treatments more efficient could be greatly simplified by an understanding of the specific change processes that make the treatments work. But a close review of child research reveals much more about what outcomes are produced than about what actually causes the outcomes." Similar statements have been made about moderators of treatment effects, at least in part due to criticisms that one treatment cannot be ideal for all persons. As stated by Kraemer, Frank, and Kupfer (2006, p. 2011), "In studies in which moderators are ignored, effect sizes may be biased, power attenuated, and clinical important information overlooked. Given the low costs of such analyses relative to the costs of the RCT and the possible clinical, research, and policy importance of such findings, such approaches are at least worth serious consideration."

The promise of mediation analysis in treatment research is that the analysis identifies underlying mechanisms by which treatment actions lead to beneficial outcomes. Identifying these mechanisms leads to improvements in treatment by providing clear targets for emphasis. Furthermore, if the mechanisms identified are fundamental to clinical behavior change, the mechanisms contributing to one outcome in one population may generalize to other outcomes in other populations as well. The promise of moderating variables in treatment research is that the most effective treatments for specific groups of individuals can be identified, maximizing overall treatment by tailoring treatment content to these groups.

Clinical treatment processes are complex. Consider, for example, theories explaining the etiology of depression: they include psychosocial models that postulate psychological and interpersonal causes of depression; psychoanalytic models that suggest intrapsychic influences; behavioral models that emphasize learning, habit, and environmental causes; cognitive models that emphasize perceptual and attributional styles that underlie depression; biochemical models that postulate chemical imbalances for the cause of depression; and genetic models that implicate genes or gene–environment interactions as the cause of depression. Furthermore, the process of change may be a chain of events with components of different theories operating in different parts of these changes such as the chain that connects negative life events to hopelessness to learned helplessness to depression (Kazdin, 1989). Different theoretical models postulate different mechanisms for etiology and corresponding treatment mechanisms to activate for change. And many possible mechanisms may be at work in a treatment, as discussed by Freedheim and Russ (1992), who identified six mechanisms of change in child psychotherapy: (1) correcting emotional experience so that the child's emotions are valid, (2) insight into emotional aspects of conflict and trauma, (3) labeling of feelings to make them less overwhelming, (4) development of skills and coping strategies to solve problems, (5) exposure to a consistent, predictable, and caring therapist, and (6) nonspecific factors such as expectations of therapy or belief in the success of therapy prior to therapy, the therapeutic alliance between therapist and client, and a host of mechanisms related to compliance with prescription drug regimens. Although a treatment may be designed to be specific to a disorder, client populations may be heterogeneous on a variety of factors, including demographic and personality characteristics and comorbidity, which may lead to differential levels of treatment efficacy for different subgroups.

There are several ways that mediation and moderation analysis in the context of clinical treatment research differ from applications in other contexts, such as prevention research. One difference, especially in the context of psychotherapy research, is that there are many possible levels of treatment intervention. Psychotherapy may be delivered in a group or family setting, as well as one on one. There are several agents of change in psychotherapy. The psychotherapist may perform activities designed to be therapeutic, such as interpreting a client's statements or assigning homework. The client responds to the therapist's interventions with thoughts and actions that activate internal change mechanisms. Clinical treatment may also include environmental changes such as inpatient hospitalization. Drug treatments may also be part of a treatment program and may be changed during therapy. The many different agents of change may work simultaneously or synergistically in a treatment program. The meaning and effectiveness of each action may differ based on the individual characteristics of the client. As a result, clinical research on mediating and moderating variables can be complicated, involving potentially complex theories about how treatments work,

including how they work over time, and for whom they work. There are also extensive practical research design issues of what mediators and moderators to measure, when to measure mediators and outcomes, and how long effects are expected to occur. These are very challenging issues that can be addressed by incorporating a wide variety of information to identify mediation and moderation. It is unlikely that one study or type of design would be sufficient to thoroughly investigate mediating and moderating processes in treatment research. As a result, the general approach of this chapter is to acknowledge that identifying mediating and moderating variables requires information from many sources, including clinical judgment, qualitative information, as well as carefully controlled studies to investigate how and for whom treatments are effective.

This chapter outlines the methods for identifying the mediating mechanisms that lead to differential treatment effects and the methods for identifying moderators of these effects, and provides citations for the reader to learn more about these methods. First, we describe several types of third variables, including the mediating and moderating variables that are the focus of this chapter. Second, we describe reasons for conducting mediation analysis in treatment research and provide examples of theoretical mechanisms in several areas of treatment research. Examples from several areas are selected because they illustrate mediation and moderation for treatments that differ by age, use of medications, and mediating variable targets. The statistical mediation model is then described. Third, we describe reasons for moderation analysis of treatment studies with examples and the simplest statistical moderation model. Fourth, we outline design issues for mediation and moderation analysis of treatment studies. Fifth, we describe modern statistical developments for mediation and moderation models. Sixth, we outline an exemplar treatment mediation and moderation study including multiple mediators and longitudinal data. Finally, we describe advanced models that may more accurately reflect the complexity of mediation and moderation analysis in clinical treatment and suggest several future directions for the assessment of mediation and moderation in treatment research.

Definitions of Third Variables

Mediators and moderators are examples of third variables—that is, variables with effects that clarify or elaborate the relation between an independent variable and a dependent variable (MacKinnon, 2008). Consider two variables: *X*, coding whether a participant received a new treatment or a standard treatment, and *Y*, an outcome variable such as depression. Treatment studies may include a third variable hypothesized to help explain how the new treatment would outperform the standard treatment (that is, a mediator variable), or a third variable hypothesized to help determine if the relative performance of the new and standard treatments varied for different client subgroups or treatment contexts (that is, a moderator variable). *Third variables* such as these are useful for clinical research because they have the potential to provide a more detailed description of the relation between treatment and outcome, ultimately informing the design and application of clinical interventions. Four major types of third variables are common in clinical research models. A *mediating variable* represents the intermediate member of a causal chain of relations, such that *X* causes *M*, the mediator variable, and *M* causes *Y*. A statistically equivalent, although conceptually distinct, third variable is a *confounder* (MacKinnon, Krull, & Lockwood, 2000), in which a third variable, *Z*, covaries with both *X* and *Y*. The difference between a confounder and a mediator is that a confounder is *not* part of a causal sequence. Because confounders and mediators cannot be distinguished with statistical methods, it is important that researchers have a clear theoretical basis for choosing a mediation model, and where possible, that they design a study that maximizes the potential for causal inference by including measures of important confounding variables.

Third variables also include *moderators* and *covariates*. A moderator affects the strength of a relation between two variables, such that the strength of the relation is dependent on the value of the moderating variable, *Z*. Moderating influences are particularly important for treatment studies because they yield information about which groups can most or least benefit from a given clinical strategy or under what factors or conditions a particular treatment yields the most benefit. For example, a randomized trial comparing individual therapy and couples therapy for women with alcohol use disorder showed that participants who had DSM-IV Axis I disorders had a greater percentage of days abstaining from alcohol when they were exposed to the couples therapy condition than individual therapy (McCrady, Epstein, Cook, Jensen, & Hildebrandt, 2009). Thus, the relation between the intervention condition (*X*) and percentage of days abstinent (*Y*) depends on the presence of an Axis I disorder (*Z*).

Finally, covariates are third variables that can improve the ability of *X* to predict *Y* because these variables parse the variance of *Y*, such that *X* predicts only the part of *Y* that is not predicted by a covariate, *Z*. Covariates are related to *Y* and not substantially related to *X*, so they do not appreciably change the *X*-to-*Y* relation. Covariates are a fundamental feature of analyses of treatment effects. At a minimum, the baseline outcome measure should be included as a covariate, thereby increasing statistical power and interpretation of results in terms of change. The inclusion of additional covariates (and confounders) is driven by theoretical and empirical considerations for the population and treatment goals under study. In the context of treatment, a covariate may be a pretreatment attribute that predicts the outcome similarly across all treatments, and if significant, the pretreatment attribute is called a *prognostic indicator* because it gives the same relative prognosis to subgroup members regardless of what treatment they are exposed to.

Examples of Moderating Variables

There are a large number of potential moderating variables because a treatment may have effects that depend on many different personal characteristics or on many different factors in the administration of treatment. In the moderation model examining subgroup characteristics, a pretreatment attribute interacts with the treatment variable to predict outcome, such that the attribute's subgroups have a different treatment effect. Therefore, moderation is sometimes referred to as an "attribute-by-treatment interaction." The pretreatment attribute (moderator) in a treatment moderation model is sometimes referred to as a "prescriptive indicator," because the differential response represented by the interaction suggests that treatment effects on the outcome would be optimized by "prescribing" different treatments for different subgroups (matching subgroup to treatment). For example, before a tobacco cessation study, clients may be in different stages of smoking cessation (precessation, cessation, and maintenance), so the effect of a treatment likely will differ across this moderator. Sotsky, Glass, Shea, and Pilkonis (1991), in an exploratory analysis of data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (TDCRP; Elkin, Parloff, Hadley & Autry, 1985), grouped potential treatment predictors (including both prognostic and prescriptive indicators) into three domains that can be generally applied in clinical research: (1) sociodemographic variables, (2) diagnostic and course variables, (3) and function, personality, and symptom variables. In several research studies, function, personality, and symptom variables have been identified as (potential) moderators of depression or alcohol dependence treatment. In a study comparing cognitive-behavioral therapy (CBT) to antidepressant medication plus clinical management (ADM) for severely depressed outpatients (DeRubeis et al., 2005), personality disorder status was found to differentially predict acute treatment response such that the absence of a personality disorder predicted better response to CBT, and having a personality disorder predicted better response for ADM (Fournier et al., 2008). In a post hoc reanalysis of TDCRP data, personality disorder was found to differentially predict acute response to CBT and interpersonal therapy (IPT), such that CBT was more effective for clients with avoidant personality disorder and IPT was conversely more effective for clients with obsessive-compulsive personality disorder (Barber & Muenz, 1996). In a study designed to assess moderation hypotheses in the treatment of alcohol dependence (Project MATCH; Longabaugh & Writz, 2001), a pretreatment measure of patient anger moderated the treatment response of outpatients at 1- and 3-year follow-up, such that patients higher in anger responded better to motivational enhancement therapy compared to CBT and twelve-step facilitation, with the reverse true for patients lower in anger.

Although moderation often refers to a situation in which pretreatment characteristics interact with treatment to predict outcome, and thus may address questions pertaining to which treatment works for whom, an interesting exception is when treatment moderates the relationship between a during-treatment variable and outcome. In a hypothetical example presented by Kraemer and colleagues (2008), treatment moderates the relation between experiencing during-treatment traumatic events and a coping outcome. In a study comparing CBT to ADM for depressed outpatients, DeRubeis and colleagues (1990) found, in the absence of a treatment effect on either depression or early dysfunctional attitude change, that treatment interacted with early dysfunctional attitude change to predict subsequent outcome; early dysfunctional attitude change predicted subsequent outcome in the CBT group but not in the ADM group. In cases like this, in which a during-treatment variable is a potential treatment mechanism, the interaction is suggestive of differential mechanisms; one explanation for the finding is that CBT improves depression

by changing dysfunctional attitudes, whereas ADM improves depression by some other means.

Examples of Mediating Variables

Although clinical treatments are often introduced based on hunches and common sense, scientific clinical research focuses on the theoretical bases for treatments and evidence for treatment effectiveness (Kazdin & Nock, 2003). As a result, modern clinical treatments have a theory for the mechanisms by which treatment is believed to cause change in an outcome variable. In tobacco cessation, for example, craving is often a primary target of intervention based on the theory that a major cause of lapse is feelings of craving (Baker et al., 2011). In the case of smoking, there are several additional theoretical perspectives with different corresponding mediating variables. Self-determination theory focuses on the client's self-efficacy and intrinsic motivation to quit (e.g., Williams et al., 2006). In addition to norms for smoking, social cognitive theory targets the client's confidence to quit smoking, self-efficacy to resist smoking in social situations, self-efficacy to resist smoking in stressful situations, and response to lapses (e.g., Bricker et al., 2010). Coping theory targets the client's craving, anhedonia, ability to cope with cues that tend to trigger smoking, response to lapses, and social support (e.g., Shiffman, 1984). Negative reinforcement theory focuses on withdrawal symptoms (e.g., bupropion literature; Piper et al., 2008). Potential mediating mechanisms for alcohol addiction are equally interesting, such as client change talk as a mediator of therapist motivational interviewing behaviors and outcome (Moyers, Martin, Houck, Christopher, & Tonigan, 2009). Although not universally considered a clinical treatment, Alcoholics Anonymous has clearly defined mediating targets thought to cause abstinence, such as spirituality as a mediator of recovery (Kelly, Stout, Magill, Tonigan, & Pagano, 2011).

Similarly, different psychosocial treatments for major depressive disorder are based on different theoretical mechanisms. In some psychodynamic approaches (e.g., Luborsky et al., 1995), the therapist offers interpretations of the client's statements intended to identify the client's core conflictual relationship patterns in order to improve the client's understanding of these patterns (Gibbons et al., 2009). In behavioral therapy for depression, therapists use techniques that encourage clients to reduce avoidant behavior and increase rewarding activities (Dimidjian et al., 2006). According to cognitive approaches (e.g., Beck, Rush, Shaw, & Emery, 1979), negative cognitive processes, including cognitive content and cognitive style, are responsible for maintaining depression and are targeted for change. As such, the therapist teaches the client to independently identify and evaluate evidence for and against depressotypic thoughts so that the client will be able to do so outside of therapy and thus remain nondepressed (i.e., acquire "compensatory skills"; Barber & DeRubeis, 1989) or experience a change in underlying cognitive schema and thus be less likely to have these types of thoughts in the future.

In autism treatment for children, the primary goal is often to reduce the effects of autism on the child's functioning rather than to achieve remission or recovery, and many interventions are designed to act through parenting mechanisms as well as the child's behaviors. For example, there are treatment programs designed to improve parental understanding of autism; increase parental warmth, sensitivity, and responsiveness; and reduce parenting-related stress. Changing these variables is hypothesized to lead to better parent–child interaction styles, which in turn leads to improved child behavior (Happé & Ronald, 2008; Ratajczak, 2011).

Reasons for Mediation Analysis in Treatment Research

Mediating variables are useful in treatment research based on a variety of overlapping reasons (see also MacKinnon, 1994, 2008, 2011; MacKinnon & Luecken, 2011). One reason for the inclusion of mediation variables is for a *manipulation check.* The use of mediation analyses provides a way to confirm that varying treatments produce varying levels of change in the hypothesized mediating variables (i.e., the treatment manipulation should theoretically produce more change in the mediator than does the control manipulation). For example, consider a treatment that was designed to increase clients' confidence to quit smoking (the mediator), which is, in turn, hypothesized to increase rates of smoking cessation (the outcome). In this scenario, a greater effect on confidence to quit smoking should be observed for those in the treatment condition than those in the control condition. If the treatment does not have beneficial effects on the mediator, then it is unlikely that the treatment will affect the outcome even if the mediator is causally related to the outcome. However, failure to obtain a significant treatment effect on hypothesized mediating variables does not necessarily mean that such a relationship does not exist, owing to chance, for example.

Another reason to include mediation analyses in treatment research is for *treatment improvement*. Mediation analysis can generate information to identify successful and unsuccessful portions of a treatment. Additive studies, in which a new component is added to an existing protocol, and dismantling studies, in which individual components are examined separately or in combination, enable researchers to evaluate individual treatment components (Resick et al., 2008). In a smoking example where craving for tobacco is hypothesized to be the mediator and smoking cessation is the desired outcome, if the treatment does not influence the measure of cravings for tobacco more than the control condition, then that treatment may need to be reworked. In some cases, individual treatment components can be evaluated such that if the component does not change a proposed mediator, then that particular component can be targeted for improvement or omission from treatment. Suppose that a requirement for daily journaling was included in a smoking cessation treatment based on the theory that journaling would reduce cravings. By providing versions of the treatment with and without the journaling requirement, researchers can understand whether or not the journal requirement contributes to changes in craving. Additionally, mediation analysis can identify "faulty" mediators. If the treatment significantly reduces cravings to smoke compared to the control condition but does not differentially affect the outcome, this provides evidence that craving reduction may not be a mechanism for smoking cessation. New mechanisms would have to be proposed, with treatments designed to target them. Related to identifying faulty mediators, mediation analysis may also help identify counterproductive mediation processes by which treatment affects mediators in a way that there are iatrogenic effects on the outcome (MacKinnon et al., 2000).

Measurement improvement may also be a benefit of mediation analysis. Lack of a treatment effect on a known mediating variable may suggest that the measures of the mediator were not reliable or valid enough to detect differences. As an example, if no treatment effects are found on confidence to quit smoking, it may be that the method used to measure confidence is not as reliable or valid as needed. For this reason, it is possible that mediation analysis may lead to a more refined measure of a particular aspect of a general mediator.

As another reason supporting the inclusion of mediation variables, mediators enable research to be conducted that allows for the *possibility of delayed effects*. There are many examples of research where the expected treatment effect on an outcome is not expected to occur until later in time. For example, the effects of a differential increase in perceived self-efficacy to resist smoking in stressful situations may not be apparent at the end of the acute treatment period, but may emerge during a follow-up period when there have been more opportunities for exposure to stressful life situations that tend to precipitate smoking relapses. Thus, initial changes in the mediator during the acute treatment period may be indicative of a successful delayed outcome.

Mediators may also provide additional theoretical understanding of what makes one treatment condition work better than another, thus enabling researchers to *evaluate the process of change*. Mediation analysis provides information to help answer the question "what are the processes by which this treatment affects an outcome?" For example, it is possible to study whether certain processes, including particular therapist interventions, are related to treatment and outcome such that they may be mediators (see also Chapter 9 in this volume). However, in the context of flexible psychosocial treatments provided in a one-on-one setting, it is important to consider how the responsiveness of therapists to their clients' needs might influence process–outcome correlations (Stiles & Shapiro, 1994). That is, in contrast to a standard treatment mechanism hypothesis in which the delivery of an effective treatment component would be positively and uniformly associated with better outcome, the "responsiveness critique" (Doss, 2004) suggests that the treatment provider adjusts the level of component delivery to meet the needs of the client. Because more impaired and potentially more difficult clients may require more of the component, there may be a zero or negative relation between delivery of the component and client outcome for given individuals, even if the component is effective. This pattern of results would be an example of moderation of a mediated effect, where the mediating process may differ across groups of people.

One of the greatest strengths of including mediating variables is the ability to test the theories upon which treatments were based, including the ability to examine competing theories. Thus, one of the reasons for including mediating variables is for *building and refining theory.* Mediation analysis in the context of randomized control trials is optimal for testing theories as well as for comparing outcomes. Competing theories for smoking cessation, for example, may suggest alternative theoretical

mediating variables that can be tested in an experimental design.

There are many practical implications for use of mediation analyses in treatment research. For instance, if there are many components included in a treatment, mediation analyses can help determine which components are crucial to the desired outcome. In the multicomponent CBT for child anxiety, changes in negative self-talk were identified as mediators of treatment-produced gains (Kendall & Treadwell, 2007; Treadwell & Kendall, 1996), suggesting that this component merits being included within the treatment (along with exposure tasks). Determining which pieces of treatment are necessary becomes particularly important when there are limited resources available both for research and for the actual implementation of treatment programs. Treatment programs will cost less and provide greater benefits if the critical ingredients can be identified, the critical components retained, and the ineffective components removed. Additionally, mediation analyses can help researchers decide whether to continue or discontinue a relatively unsuccessful treatment approach by providing information about whether the lack of success was due to a failure of the intervention to change a hypothesized mediator (action theory), a failure of the hypothesized mediator to significantly predict outcome within each treatment (conceptual theory), or a failure of both action and conceptual theory.

Reasons for Moderation Analysis in Treatment Research

There are many reasons for moderator variables in treatment research as for other research projects (Mackinnon, 2011). The use of moderators *acknowledges the complexity of behavior* by focusing on individual differences and personalized therapeutic approaches to help answer the question "for whom does this treatment work?"

Testing moderation effects can assess *specificity of effects* by identifying groups for which a treatment works best or groups for which the treatment doesn't work at all, or groups that experience iatrogenic effects. An understanding of how the treatment works for various subgroups could then be used to tailor treatments for particular subgroups. For example, many psychiatric conditions have high rates of comorbidity (e.g., depression and anxiety; Pini et al., 1997). It may be the case that a particular treatment or intervention works (or doesn't work) only for those clients with (or without) a particular comorbidity. Although potential variables accounting for treatment differences among subjects are numerous, a moderator does not necessarily have to test a subgroup of people. In general, a moderator may provide an answer to the question "under what conditions or factors does this treatment work?" There may be attributes of the therapist, location (inpatient vs. outpatient), or context in which the patient functions (e.g., type of family, support system; Kazdin, 2002) that may influence a treatment effect. Clinical psychology is abundant with potential factors that may moderate treatment effects. By understanding these moderating factors, treatment can be tailored to maximize the factors that contribute to better outcomes.

In contrast to specificity, moderation analysis allows for *generalizability of results.* A test of moderation can provide insight regarding whether a treatment has similar effects across all groups. In this case, assuming a sufficiently large sample size, a nonsignificant interaction may tell us that a treatment is not moderated by a particular subgroup (e.g., males vs. females) and thus would be an appropriate treatment to apply to both groups. The consistency of a treatment across subgroups (or other factors) demonstrates important information about the generalizability of a therapeutic technique.

Another reason moderation analysis is useful in treatment research is that it may *identify iatrogenic effects*. The use of moderation can identify subgroups (or factors) for which effects are counterproductive. There may be cases where the main effect of treatment across a diverse population has a positive change on the outcome variable of interest, but it is possible that there are subgroups within the diverse population for which the therapy actually has the reverse impact on the outcome. Returning to the comorbidity example, it could be that a particular treatment causes significant improvement in clients with depression without comorbid anxiety, but the same treatment may actually worsen anxiety symptoms in those who have both depression and anxiety.

Moderation analyses may also be used as a *manipulation check*. For example, if a particular therapeutic intervention is proposed to work better after greater exposure to the therapy, then manipulation of the number of hours of treatment ("dosage") may serve as a moderator. If the treatment is successful, the size of the effect should differ across the dosage of treatment, assuming monotonic effects of treatment. Similarly, there may be research contexts in which the test of moderation is a test of theory, such as when treatment effects are expected

for one group but not another. For example, depression treatment effects may be strongest for persons severely depressed compared to persons with mild depression because the content of the therapy may more directly target the severely depressed group. Conversely, greater treatment effects may be observed for persons with mild depression because severe depression may interfere with accomplishing the treatment tasks.

An important use of moderation analysis is to *investigate the lack of a treatment effect*. If two groups are affected by treatment in an opposite way, the statistical test of the overall effect may be nonsignificant even though there may exist a statistically significant treatment effect for both groups, but in opposite directions. To observe these effects, the investigation of moderating variables is needed. Without understanding that such subgroups exist, treatment that is ideally suited for a particular subgroup may be deemed ineffective and abandoned. Also, moderation analysis can be used to examine whether two similarly effective treatments are "differently better" for different subgroups.

The practical implications of finding (or not finding) moderator effects are considerable. For example, if a therapy produces positive results for all levels of a moderator, it is reasonable to conclude that the therapy is effective across these subgroups or factors. However, if a moderator effect suggests that one or more treatments work well in particular subgroups or in particular contexts but not others, this should inform for whom and under what conditions the particular treatment is given. This would also suggest that other treatments need to be developed or used to address the situations in which the treatment did not work as well. An understanding of the characteristics that affect the effectiveness of treatments across groups can potentially lead to better overall treatment outcomes.

A Framework of Influences in Clinical Mediating Mechanisms

There are several unique aspects of investigating mediation mechanisms in treatment research. As described earlier, treatment may be delivered in groups, families, or one-on-one settings. The primary agent of change in therapy is typically the client, who processes therapeutic interventions from many different levels. In other cases, the agent of change may be the couple or the family, for example as targeted by marital or family therapy. The most recognized source for change is the therapist, who conducts actions designed to assist the client. Activities within the client include his or her own actions and thoughts. In addition, clinical treatment may also include environmental changes designed to enhance treatment, such as a period of separation as part of marriage therapy or hospitalization to focus on specific behaviors. In addition, drug treatments may be included and possibly changed during treatment. These different agents of change, therapist, environment, and client, may work simultaneously or synergistically in a treatment program. As a result, clinical research requires rather detailed development of theory relating treatment actions to mediators and also theory for how the targeted mediators are presumed to be related to outcome variables. A further complication for mediation models is that treatment is often adaptive, so the meaning of the client's experience of actions and the mediating processes themselves may differ at different times for the client. Because therapy is typically conducted when a problem already exists, treatment is change-focused by its nature. In comparison, prevention-focused mediation often addresses methods to maintain ongoing positive behavior to prevent negative behavior from starting. In treatment, study participants start with negative behavior and the goal is to change often-habitual negative behavior.

A general treatment mediation model has these two main components, which describe the underlying theory of the mediation process (Chen, 1990; MacKinnon, 2008). Figure 15.1 shows a basic treatment mediation model, in which the component containing the *a* path represents *action theory* and the component containing the *b* path is labeled *conceptual theory.* The action theory of the model is the point of the mediation process that the researcher "acts" upon. If the action theory of the treatment is supported, then the researcher has designed and delivered an effective treatment, such that the mediator changed in the desired manner for the participants in the treatment group. Statistically,

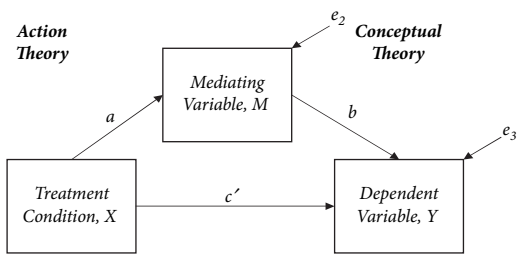


Figure 15.1 Single-mediator model showing action and conceptual theory components.

this scenario would result in a significant *a* path in the expected direction. The conceptual theory component of the model is *not* directly within the control of the researcher to produce change (see, however, the section on experimental approaches to mediation later in this chapter for ways to potentially overcome this limitation). The term "conceptual" is used because the researcher must construct a hypothesis based on substantive theory and prior empirical evidence instead of manipulation in a controlled setting. If the conceptual theory component of a mediation model is supported, then the relation between the mediator and the outcome after accounting for the effect of treatment condition occurs in the expected direction (significant *b* path). Conceptual theory is what is typically referred to by theory in treatment research, and the mediators selected for target in treatment research ideally have substantial prior research evidence for their role in causing the outcome variable.

Constructing a treatment mediation hypothesis requires thoughtful consideration of the unique challenges of each component of the model. Good conceptual theory ensures that the mediators are theoretically and/or empirically linked to the outcome of interest. Good action theory ensures that a treatment will change the mediators identified in the conceptual theory in the desired way. Thus, identifying and measuring the "right" mediators is a critical task in the development of treatment studies to assess mediating mechanisms. Given the importance of action and conceptual theory, how can a researcher decide which mediators to include in a treatment study? Although it is impossible to specify a one-size-fits-all approach for every case, two requirements of treatment mediators must be satisfied: (1) the mediators must be modifiable and (2) the mediators must be expected to precede the outcome (i.e., temporal precedence). Of course, there are additional considerations depending on the outcome and population under examination. We outline these considerations below that follow directly from the specification of action and conceptual theory. *What types of mediators can be reasonably expected to change, given the time and resources allotted to the study?* Mediators in treatment studies can be classified in two major categories: (1) external mediators, which include environmental variables that are "outside" of a person, and (2) internal mediators, which refer to mediating variables representing phenomena occurring only within or by that person. External mediators in treatment studies can be social (e.g., exposure to deviant peers) or physical (e.g., access to junk food). Internal mediators can represent cognitive (e.g., rumination) or behavioral (e.g., social skills, sleep) phenomena. Careful consideration of the types of mediators available is important because attempting to change internal versus external mediators will often demand different strategies for designing treatment conditions. For example, a treatment designed to reduce adolescent problem behavior could be constructed to target youths' social environments by reducing exposure to deviant peers (an external mediator). Another treatment component to reduce youths' exposure to deviant peers may include occupying the youths' time with constructive activities away from their usual peer network. Alternatively, a treatment for adolescent problem behavior could target *internal* mediators such as youths' social competence, which may be expected to predict problem behavior. Clearly, designing treatment conditions with a goal of altering the youths' external environments is a different task than attempting to change an internal characteristic such as social competence. Although researchers can, and often do, target both internal and external mediators, treatment can become overly complicated if the researcher attempts to target a wide range of both internal and external mediators. The methodological challenge of mediation in treatment research revolves around the measurement of the many treatment components and targets on clients. Although the challenge of teasing apart the many different program components is daunting, the explicit theoretical specification of these types of effects based on theory prior to a treatment study is ideal for treatment science.

Statistical Analysis of the Single-Mediator Model

In the single-mediator model, mediated effects can be estimated based on information in three regression equations, although only two of the three equations are necessary to test for mediation (Mackinnon, 2008). The simplest mediation model is described first to clarify ideas; the single-mediator model comprising one independent variable, *X*, one mediator, *M*, and one outcome variable, *Y*, is visually illustrated by the set of path diagrams shown in Figure 15.1. The mediation model is a causal model such that *X* causes *M* and *M* causes *Y*. Thus, *M* and *Y* in these equations could also be considered as change in *M* and change in *Y* corresponding to two waves of longitudinal data with baseline measured before treatment is delivered. To probe more deeply into this model, we will consider the basic

single-mediator model in terms of a substantive example, although the actual model used in treatment research is likely to include multiple mediators and longitudinal measures. The single-mediator model is described as this model illustrates the complexity of identifying mediating mechanisms even for the simplest mediation model. Consider research by Kaskutas, Bond, and Humphreys (2002) that examines whether the relationship between Alcoholics Anonymous (AA) involvement and reduced substance abuse is mediated by changes in social networks. For purposes of this example, we will assume that social network and substance abuse are measured as single variables.

For the first equation, the dependent variable is regressed on the independent variable.

$$Y = i_1 + cX + e_1. \quad (1)$$

Equation (1) represents a simple relationship between the independent variable, *X*, and the dependent variable, *Y*, corresponding to whether treatment had an effect on the outcome. In terms of the example, Equation (1) determines if there is an association between AA involvement (*X*) and substance abuse (*Y*) represented by the *c* coefficient. The intercept of the equation is *i* and the residual variance (i.e., the part of the criterion not explained by the relationship with *X*) is represented as e_1.

In the second equation, the dependent variable is regressed on both the independent variable and the mediating variable.

$$Y = i_2 + c'X + bM + e_2. \quad (2)$$

Equation (2) answers the question: "Can we predict substance abuse from both participation in AA and social network?" In this equation, *b* and *c*′ are partial regression coefficients, each representing the effect of the predictor on the outcome controlling (or adjusting) for the effect of the other predictor. While *c* reflects the unelaborated relation between *X* and *Y*, *c*′ represents the relation between *X* and *Y* after adjusting for the mediator. Thus, *c*′ can be used to address the question "After adjusting for the impact of social networks on substance abuse, what is the remaining relationship between AA involvement and substance abuse?" The i_2 and e_2 terms represent the intercepts and residual variances, respectively.

In the third equation, the mediator is regressed on the independent variable.

$$M = i_3 + aX + e_3. \quad (3)$$

This regression equation addresses whether there is a relationship between AA involvement and social network, as represented by the coefficient *a*. In this equation, the i_3 term is the intercept and the e_3 term is the residual variance. Notice that equations (2) and (3) both include the mediating variable; in combination, these equations represent the mediation component of the model.

The four coefficients (i.e., *a*, *b*, *c* and *c*') resulting from equations (1) through (3) can be used to quantify the mediated effect. The product of the *a* and *b* coefficients, *ab*, is the mediated effect (sometimes called the indirect effect). The mediated effect is also equal to the total effect minus the direct effect ($c - c'$). These two measures of mediation are numerically equivalent for linear models without missing data but not for nonlinear models such as logistic regression (MacKinnon & Dwyer, 1993). The product of coefficients method is the most general method to estimate mediation applicable to simple and complex models. The most accurate methods to test the statistical significance of the mediated effect and construct confidence intervals use methods that accommodate the distribution of the product either using resampling methods such as bootstrapping or information on the critical values of the distribution of the product (MacKinnon, Lockwood, & Williams, 2004). Many computer programs now include bootstrap sampling, including Mplus (Muthén & Muthén, 2010), or methods based on the distribution of the product (MacKinnon, Fritz, Williams, & Lockwood, 2007; Tofighi & MacKinnon, 2011). More detail on these different tests can be found in MacKinnon (2008).

Despite limitations, the most widely used method to test for mediation is the causal steps method, which originated with Hyman (1955), Judd and Kenny (1981), and Baron and Kenny (1986). The causal steps method consists of a series of significance tests based on equations (1), (2), and (3) that indicate whether a variable functions as a mediator. There are substantial shortcomings of the method, such as low statistical power, and better methods are available based on resampling techniques or methods based on the distribution of the product (MacKinnon, Lockwood, & Williams, 2004). In particular, the first step in the causal steps method requires that *X* produces a change in *Y* as represented by equation (1), but mediation can exist whether or not *X* is significantly related to *Y* (MacKinnon et al., 2000). Nevertheless, the statistical significance of *X* on *Y* is important for many other reasons, such as whether there is a

treatment effect. The next two steps in the causal steps method are consistent with a test for mediation called the joint significance test, which requires that there is a significant relation between *X* and *M* and a significant relation between *X* and *M* with *X* partialled out (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). These two tests are the critical tests for mediation and can be applied in many situations. However, there are times when the joint significance test is not the main mediation test of interest, such as when there are multiple mediators and the researcher is interested in the total effect of the treatment on the outcome through all mediators. As a result, the best test for mediation tests the *X*-to-*M* path and the *M*-to-*Y* path adjusted for *X*. An estimate of this effect can be obtained with the mediated effect, *ab*, along with confidence limits obtained using the distribution of the product or a resampling method such as bootstrapping. A useful aspect of the description of the causal steps is evidence for complete mediation when the effect of *X* on *Y*, after controlling for *M*, *c'*, is not significantly different from zero. If *c'* remains statistically significant, there is evidence for partial mediation.

An important addition to equation (2) is the interaction between *X* and *M*, which may provide evidence for mediation such that the relation of the mediator to the outcome differs across treatment groups (Judd & Kenny, 1981; Kraemer et al., 2002; MacKinnon, 1994; Merrill, 1994). If there is not an overall effect of the intervention on the outcome, and the *XM* interaction is substantial, it may indicate that the relation between *M* and *Y* is different across groups. For example, in a mindfulness intervention, clients in the treatment group may have a smaller relation of attention to pain and experience of pain, compared to control participants who without training have a strong relation between attention to pain and experience of pain. Such *XM* interactions could also occur when there is a nonlinear relation of *M* to *Y* such that the levels reached in the treatment groups correspond to different parts of that nonlinear relation. The *XM* interaction could also occur when there is some other unmeasured mediator that differentially affects variables in the treatment and control group. Because the *XM* interaction suggests that the relation between the *X* and *Y* differs across groups, it is not consistent with a simple hypothesis that *M* is causally related to *Y*, although there may be a causal relation that differs in strength across groups.

There are several important assumptions of the single-mediator model, many of which are relevant for complex mediation models. The mediation regression models (equations (1)–(3)) must meet all the usual assumptions of regression analyses (e.g., correct functional form, no omitted influences, accurate measurement, and well-behaved residuals; Cohen, Cohen, West, & Aiken, 2003). Beyond general ordinary least squares regression assumptions, mediation analysis has its own host of assumptions. One assumption is that there exists temporal precedence among the variables. There is an implicit assumption of the ordering of variables that assumes that *X* precedes *M* precedes *Y* in time. The methodological literature has emphasized the importance of the temporal precedence assumption (e.g., Judd & Kenny, 1981; Kraemer, Kiernan, Essex, & Kupfer, 2008; Mackinnon, 1994), and longitudinal mediation models have been proposed to address temporal precedence. In reality, longitudinal tests are not always possible, and tests of mediation are done on cross-sectional data that include no information regarding temporal precedence. As a result, the temporal precedence assumption must largely be based on the theoretical relationship between the variables for some applications, with the best evidence reserved for data obtained over time.

Another important assumption of mediation that has gotten more attention recently is sequential ignorability (Imai, Keele, & Yamamoto, 2010). Sequential ignorability refers to the assumption that the relationship between two variables is unaffected by other variables (i.e., that there exist no other covariates, confounders, or mediators that would change results if they were included in the analysis). In the single-mediator model, there are two major relationships that may be influenced by outside variables. Sequential ignorability A refers the assumption that the *X*-to-*M* relationship is not affected by other variables, and sequential ignorability B refers to the assumption that the *M*-to-*Y* relationship is not affected by other variables. The ignorability assumption for the *X-to-M* relationship (sequential ignorability A) can generally be addressed by randomizing the levels of *X*, as in most treatment research. By randomizing participants to levels of *X*, all potential additional variables are theoretically equivalent between levels of *X*. The ignorability assumption for the *M-to-Y* relationship (sequential ignorability B) is more challenging to address because clients are not typically randomly assigned to *M*, but rather self-select their value of *M*. In an ideal world, we can satisfy this assumption by randomizing *M* to participants at each level of *X*. Realistically, this double randomization does not tend to be possible, although there are some examples (MacKinnon, 2008). Given the nature of typical mediators and/or the design and

ethical issues of treatment studies, participants generally self-select their value of the mediator. As a result, we can't be certain that other variables do not influence the relation of *M* to *Y*. Although this particular ignorability assumption is difficult to satisfy, the literature suggests some promising approaches to improving causal inference by addressing the bias introduced by omitted variables (e.g., Imai, Keele, & Tingley, 2010; VanderWeele, 2008, 2010). Vanderweele (2008, 2010) has formalized several useful methods to probe bias in the mediation relationship when the assumptions of sequential ignorability have been violated. Imai, Keele, and Tingley (2010) also provide a method to assess the sensitivity of results to potential confounders. Methods based on instrumental variables may also be useful, as are methods based on principal stratification and marginal structural models (Angrist & Krueger, 2001; Frangakis & Rubin, 2002; Ten Have et al., 2007; Ten Have & Joffe, 2010). Treatment researchers are encouraged to measure confounding variables that may be related to both *M* and *Y* and consider other variables that may be the true underlying mediator of treatment effects on an outcome.

A Framework of Influences in Clinical Moderating Mechanisms

A moderator is a variable that changes the magnitude or direction of the relationship of the two other variables. In a single-moderator model, a moderator changes the relationship between the independent variable, *X*, and the outcome variable, *Y*. The moderator effect is also commonly called the interactive effect to signify that the third variable interacts with the relationship between two other variables. Moderated (i.e., interactive) effects are important in understanding under which conditions or in which subgroups a particular treatment works best. As an example, it may be that attributes of the client, therapist, or location of the treatment affect the relationship between the treatment and the outcome, and, thus, a full understanding of this relationship can help practitioners to determine the conditions under which a treatment best works.

Many moderating variables are measured as a routine part of data collection (e.g., sex, socioeconomic status, and age). Comorbidity is another important moderating variable that is often measured. Other moderating variables are baseline measures of a mediator or an outcome variable. These types of moderators investigate the reasonable hypothesis that the size of a treatment effect will depend on the pretreatment values of the outcome and mediating variables under the notion that the largest treatment effects may be observed for persons with the most room for improvement before the treatment. Again, these baseline measures are included in studies with longitudinal measurements before and after treatment. Another type of moderating variable requires specialized measurement and is likely to be included in a treatment study only if theory or past research indicates that the variable is critically important. For example, tendency towards risk-taking may be an important moderating variable but would require theory for how risk-taking moderates a treatment before resources are devoted to including it in a research study.

Moderators are most commonly exogenous to the mediation process and can either be relatively stable personal or demographic characteristics such as gender or socioeconomic status, or vary over time (though not as a function of treatment), as in the case of pubertal status or geographical location. Additionally, higher-level moderators such as school or neighborhood variables can be incorporated in a mediation analysis (Bauer, Preacher, & Gil, 2006). The influence of exogenous moderators can be at one or two places in the simple mediation model: (1) the *a* path, in which the action theory is hypothesized to differ at various levels of the moderator or (2) the *b* path, in which the conceptual theory varies in strength or sign at levels of the moderator or in both paths simultaneously.

Statistical Analysis of the Single-Moderator Model

The interaction model that tests whether a variable, *Z*, is a moderator of the effect of *X* with *Y* is shown below:

$$\hat{Y} = b_0 + b_1 X + b_2 Z + b_3 XZ \qquad (4)$$

where $\hat{Y}$ is the dependent variable;

X is the independent variable;

Z is the moderator variable;

XZ is the interaction between the moderator and the independent variable; it is formed by taking the product of each observation of *X* and *Z*; and the coefficients b_1, b_2, and b_3 represent the relation of the dependent variable with the independent variable, moderator variable, and interaction, respectively.

Moderator variables may be continuous (e.g., annual salary, score on a depression inventory, duration of treatment) or categorical (e.g., gender, ethnicity, location of treatment). When both *X* and *Z* are continuous variables, it is generally advised to center the terms (i.e., deviate each observed score around the mean) before forming the interaction term (see Aiken & West, 1991, for more details).

If the *XZ* interaction is statistically significant (i.e., the coefficient b_3 is non-zero), it is often useful to explore the conditional effects (called simple slopes) where the effect of the treatment is ascertained at values of the moderator. If *Z* is sex, exploring conditional effects would involve investigating effects for males and females separately. When the moderator is continuous, probing conditional effects involves focusing on specific values of the moderator, such as at the mean and one standard deviation above and below the mean (see Aiken & West, 1991).

Moderator relationships may occur simultaneously with mediation relationships. Simultaneous occurrence of mediator and moderator relationships has been noted as the likely case in the context of treatment studies (Longabaugh & Magill, 2011), meaning that treatment mediating processes differ for different groups. Thus, many researchers advocate modeling moderator and mediator variables in the same study (Baron & Kenny, 1986; Fairchild & MacKinnon, 2009; Mackinnon, Weber, & Pentz, 1989). Fairchild and MacKinnon (2009) describe several types of effects with both moderation and mediation for the case of prevention research that are easily extended to treatment studies. One of these types of effects is *moderation of a mediated effect*, which occurs when a moderator, *Z*, affects the direction or strength of the mediator relationship. If the moderator is binary (i.e., a two-group dummy coded variable such as sex), moderator effects can be evaluated by conducting separate analyses by group. For example, if gender is hypothesized to be a moderator, mediation analyses can be conducted separately for males and females. To test for moderation, regression coefficients, *a*, *b*, *c'*, and *c* obtained in the mediation equations above, and the estimate of the mediated effect can be compared across groups using *t* tests (see MacKinnon, 2008, p. 292). These models may also be estimated simultaneously using multiple group structural equation modeling. This is a straightforward method that facilitates analysis and interpretation of whether the mediation process differs across levels of the moderator. To include continuous moderators in a mediation model, the moderators are incorporated into equations (1), (2), and (3) as interaction terms as described in MacKinnon (2008; Chapter 10). The conceptualization of these models and methods to test them has shown considerable growth over the past 10 years (see Edwards & Lambert, 2007; MacKinnon, 2008; Muller, Judd, & Yzerbyt, 2005; Preacher, Rucker, & Hayes, 2007; Tein, Sandler, MacKinnon, & Wolchik, 2004).

Multiple Mediators and Moderators

The complexities of treating a clinical disorder typically demand an approach that addresses multiple domains of functioning of the target population. For example, a program to treat problem drinking may be hypothesized to reduce opportunities to drink alcohol, increase family cohesion, and increase abstinence self-efficacy. Multicomponent treatment mechanisms such as this are readily tested with a multiple-mediator model within a

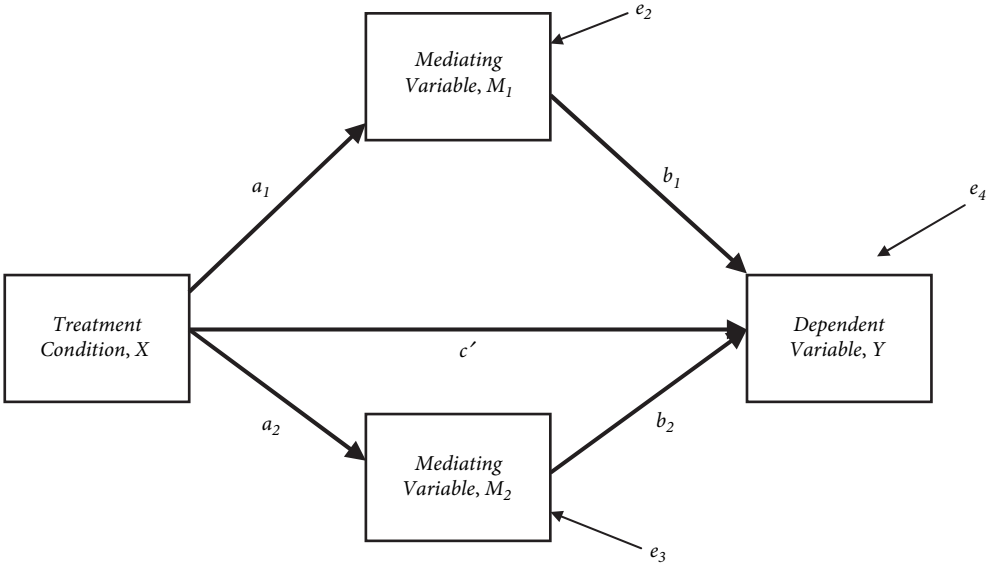


Figure 15.2 Multiple-mediator model with two mediators simultaneously operating as part of a causal process between *X* and *Y*.

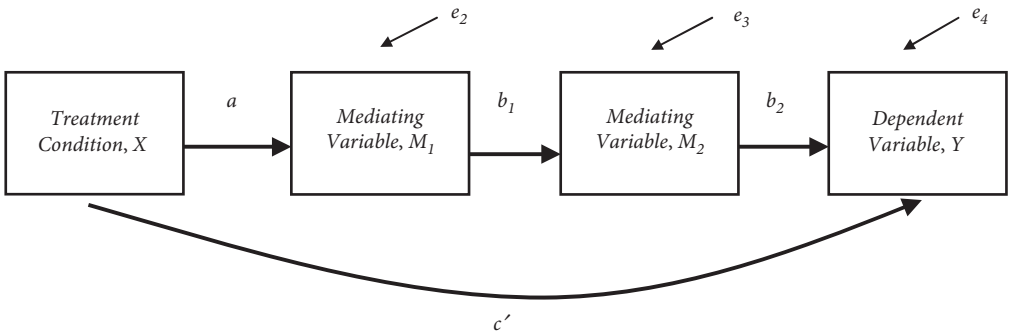


Figure 15.3 Mediation model in which a second mediator follows a first mediator in a causal sequence.

covariance structure framework such as structural equation modeling. Relations among variables can be tested in configurations in which (1) two or more mediators simultaneously operate as part of a causal process between the treatment condition *X* and outcome *Y* (Fig. 15.2); (2) two or more mediators are ordered in a sequence between *X* and *Y* (Fig. 15.3); or (3) some combination of these two configurations (Fig. 15.4).

The decision concerning the type of multiple-mediator model to use depends on action and conceptual theories. To use the problem drinking treatment example above, self-efficacy and opportunities to drink may be more proximally related to treatment condition if the treatment's action theory is supported by components to directly address these issues. Family cohesion, on the other hand, may be a more distal mediator because it involves a complex system of individuals who may or may not play an active role in the treatment. Thus, the action and conceptual theories could follow the hybrid configuration described earlier. Figure 15.5 is an example of this hybrid configuration. Most treatment studies are designed to change several mediating variables.

Tests of these models require a covariance structure software program such as SAS TCALIS, EQS

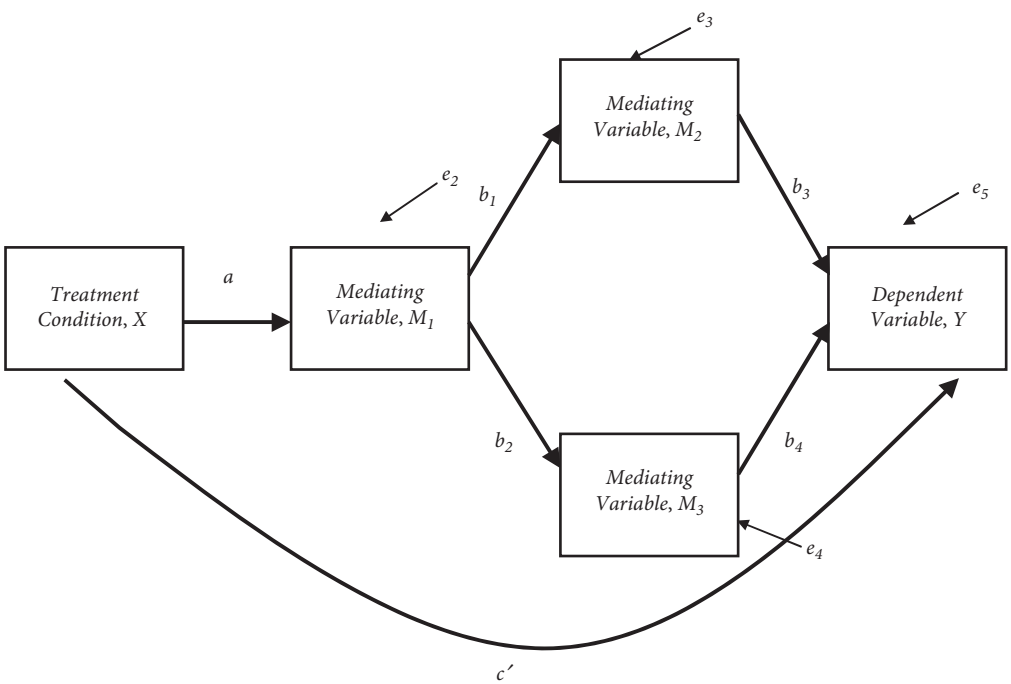


Figure 15.4 Mediation model with a combination of mediators that simultaneously operate in the mediation process and mediators that follow in a sequence.

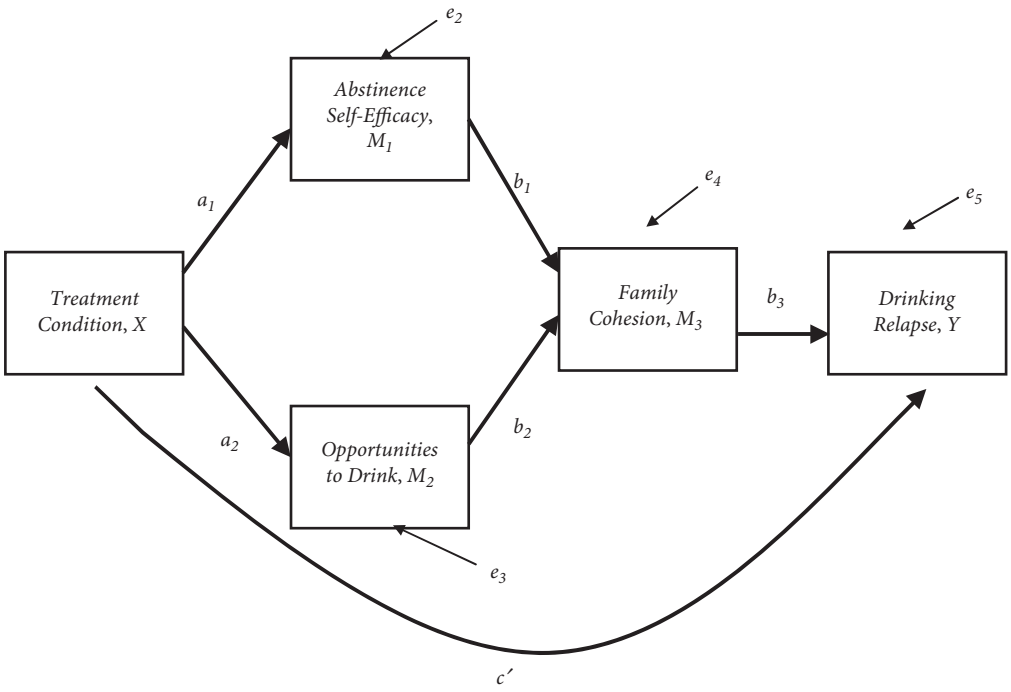


Figure 15.5 Example of a mediation model in which Abstinence Self-Efficacy and Opportunities to Drink simultaneously mediates the relation between treatment and family cohesion; family cohesion follows as a mediator in this process to predict drinking relapse.

(Bentler, 1997), LISREL (Jöreskog & Sörbom, 2001), or Mplus (Muthén & Muthén, 2010). The MODEL INDIRECT command within Mplus produces overall indirect effects as well as results for individual paths (Muthén & Muthén, 2010). Importantly, these tests can be applied using resampling methods, such as the bootstrap, that accommodate the nonnormal sampling distribution of the mediated effect.

It is possible that there are multiple moderators and multiple mediators in treatment theory and research. Testing multiple moderators is fairly straightforward: the main effect and interactive relationships would be added to equation (4). There may be main effects of each moderator and potential interactions among moderators that also may be important. For the case of multiple mediators, these variables would be added to equation (2). There would also be separate equations for each mediator, consistent with equation (3). We do not discuss the issue of multiple mediators in detail because of space limitations, but more information can be found in MacKinnon (2008).

Longitudinal Models

The mediation model is a longitudinal model such that *X* causes *M* and *M* causes *Y*. The limitations of cross-sectional measurement to assess mediation have been described, and longitudinal measurement of mediators and outcomes before, during, and after treatment is the ideal research design because change in variables can be investigated (Cheong, MacKinnon, & Khoo, 2003; Cole & Maxwell, 2003; Gollob & Reichardt, 1991; Judd & Kenny, 1981; MacKinnon, 2008). Longitudinal models best represent the temporal precedence assumption of the mediation model. Thus, assessing treatment effects requires a minimum of two measurement occasions of the mediator and the outcome. More waves of data add greatly to the information that can be gleaned from a mediation and moderation research study. Again, it is important to consider aspects of temporal precedence in these models, both theoretically before any study is conducted and then as an important part of measurement to evaluate study outcomes. Longitudinal mediation models are further complicated by the number of potential mediation pathways across time. The additional pathways require detailed thought about the timing of each variable and the timing of how the variables are related to each other. This challenge further highlights the need for models that are constructed based on theory and/or prior research. In this section, we discuss the theoretical and modeling

considerations for longitudinal mediation models of two or more waves.

Two-Wave Models

Although the use of two waves of data for mediation models has been criticized, two-wave models are relatively common in treatment research because of time and resource constraints. The two chief criticisms of these models are (1) decreased parameter accuracy of the relations among variables (Cole & Maxwell, 2003) and (2) violation of the temporal precedence assumption; either the *a* or *b* path represents a concurrently measured relation (MacKinnon, 2008). Even in the face of these limitations, however, it is possible that a two-wave model may yield accurate results (Lockhart, MacKinnon, & Ohlrich, 2011). Consider, for example, a study in which a randomized treatment condition (*X*) was hypothesized to predict adolescents' antisocial behavior (*Y*) through intentions to engage in antisocial behavior (*M*). To effectively capture the mediation process, the researcher must carefully choose the timing of measurement occasions or risk missing the mediation process altogether. Even with a three-wave (or higher) design, it is possible to miss a true mediated effect if "intentions to engage in antisocial behavior" changes after the data collection period. Alternatively, a well-thought-out two-wave design, in which the mediator and outcome are concurrently measured at wave 2, could capture the true mediation process. The decision about measurement occasion timing in the two-wave case should be based on (1) the developmental stage of the population under consideration and (2) prior evidence and/or theoretical reasoning that the conceptual theory component of the mediation model occurs in a relatively short period of time.

Autoregressive Models

For studies in which *X* represents a treatment condition, autoregressive mediation models are path models in which contemporaneous and longitudinal relations among *X* (measured only once), *M*, and *Y* (each measured three times) across three or more time points can be tested. Path coefficients indicate the stability of subjects' rank order across waves. The basic three-wave autoregressive mediation model is a path model in which relations among variables one lag (wave) apart are considered. Although it is possible to include paths that violate the ordering assumptions of mediation, this model includes only paths that make temporal sense both within and between variables. Figure 15.6 shows the basic autoregressive model for a three-wave model. The arrows labeled with *sX*, *sM*, and *sY* are the stability of the measures; a_1 and a_2 represent the two longitudinal *a* paths in the *X*-to-*M* relation; b_1 and b_2 are the two longitudinal *b* paths in the *M*-to-*Y* relations; and c'_1 and c'_2 are the two longitudinal direct effects in the *X*-to-*Y* relations. Information on other autoregressive models that include longitudinal and contemporaneous mediation can be found in MacKinnon (2008).

Although autoregressive mediation models provide a flexible way for determining the extent of stability among variable relations over time, this approach also has serious limitations (MacKinnon, 2008). For example, autoregressive models may be biased, resulting in inaccurate cross-lag coefficients (Rogosa, 1988). Additionally, because these models measure stability, or lack of a sample's movement in the levels of *M* and *Y*, they may be less useful

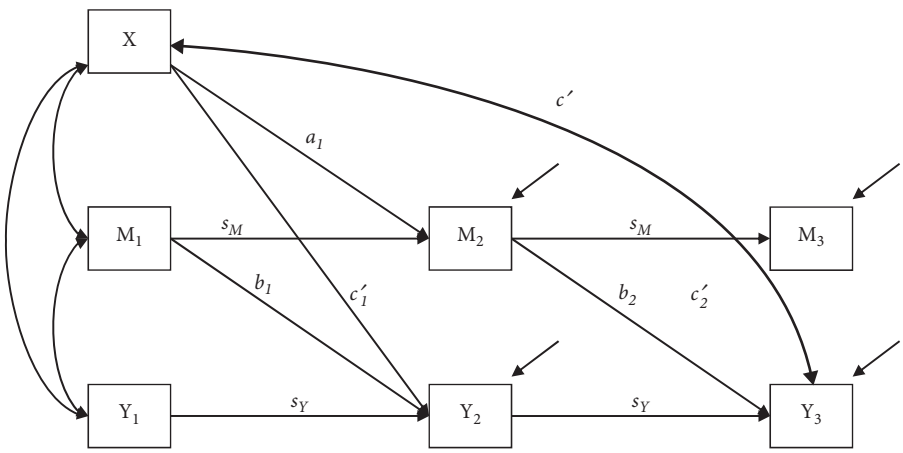


Figure 15.6 Basic autoregressive mediation model.

for models that are intended to measure significant change over time. Treatment studies with three or more waves may therefore benefit from latent growth models, which model the extent of individual change of variables and the mediation relations among these levels of change.

Latent Growth Models

Latent growth modeling is now a common method of analyzing longitudinal data. Although traditional regression and path analysis consider changes within a sample to occur at the same rate for all individuals, latent growth modeling incorporates continuous variable change (i.e., slopes) and starting points (i.e., intercept) as latent variables. Thus, change in one variable can be predicted by change in another variable. Slopes can be modeled as either linear or nonlinear. Mediation relations for treatment latent growth models are similar to those for a traditional mediation model, such that the relation between the treatment *X* and outcome *Y* is explained by both the indirect effect through the mediator and the direct effect. A critical difference is that the growth of *M* mediates the relation between the treatment *X* and the growth of *Y*.

Another longitudinal model closely related to the latent growth curve model is the latent change score model (LCM), which models differential change between two or more sets of pairs of waves (Ferrer & McArdle, 2003). This feature is attractive for treatment research because it is possible to examine whether a given treatment or treatment component was effective at specific stages of change and not others. LCMs are also readily applied to mediation models (MacKinnon, 2008), and moderation of specific paths or multiple group differences in model estimation are also possible. Figure 15.7 shows an example of a three-wave latent change score mediation model, in which *X* is a binary variable indicating treatment versus control groups. The specification of the model is flexible depending on the hypothesized relations and time-dependent expectations about change. In this example, we show a latent change mediation model in which the standards of temporal precedence reflected in the action (α) and conceptual (β) theory predict the change in *M* from wave 1 to wave 2, and this change will then predict the change in *Y* from wave 2 to wave 3. Alternatively, if a fourth wave of data were available, one could specify a relation between the change in *M* from waves 2 to 3 and the change in *Y* from waves 3 to 4 to reflect changes in intervention components. For example, a smoking intervention may introduce therapeutic components that address

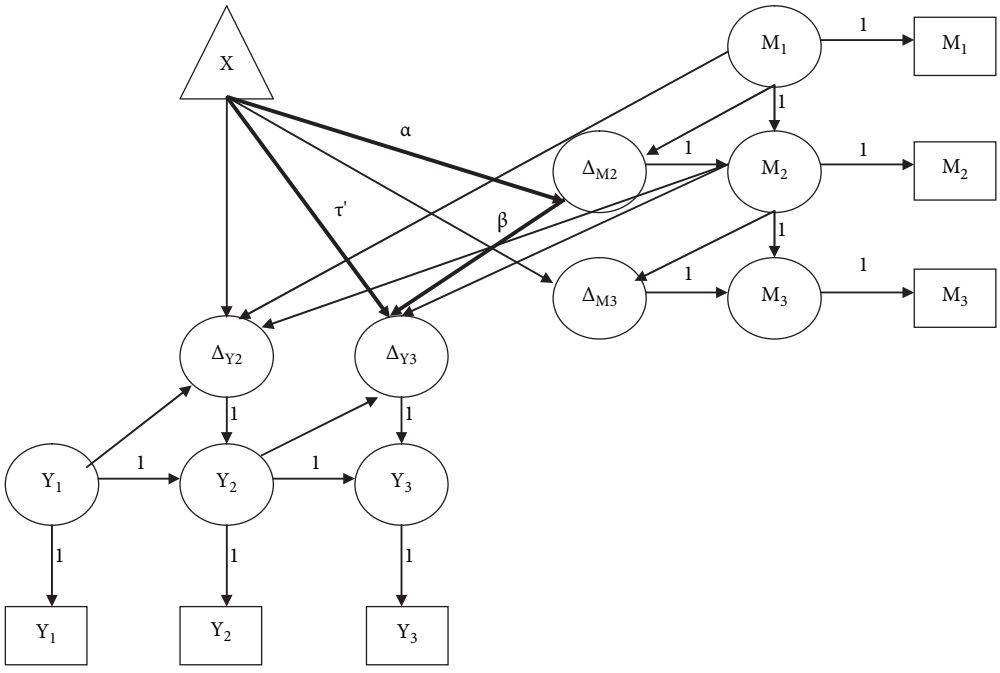


Figure 15.7 Example of a latent difference score mediation model. *X* is the treatment condition variable. Paths in bold indicate longitudinal mediation, in which *X* predicts the change in *M* between time 1 and time 2 (α), which, in turn, predicts the change in *Y* between time 2 and time 3 (β).

the maintenance phase of smoking cessation at a later time, so relevant mediators within this phase likely would not be captured until later in the assessment schedule.

Additional Longitudinal Models

There are several additional models that are relevant for treatment research on mediators and moderators. First, there are person-oriented models that focus on categorizing individuals in terms of their type of change (see Collins, Graham, & Flaherty, 1998) and configural frequency analysis (von Eye, Mun, & Mair, 2009). New mediation methods based on the analysis of data from individual participants may provide a useful way to investigate mediation processes in adaptive treatment or when few participants can be studied (Gaynor & Harris, 2008). Another important longitudinal model is survival analysis, where time to relapse is the primary dependent variable, allowing for assessment of how treatment changes a mediator that then lengthens time to relapse into drug use, to use an addiction example (Hosmer, Lemeshow, & May, 2008; VanderWeele, 2011).

Experimental Designs

In the simplest case, randomized treatment studies (randomized clinical trials) involve random assignment of participants to a treatment group (X), which allows the differences in means of groups on the mediator (M) to be attributed to the experimental manipulation. Because the mediator typically is not directly manipulated, however, the $M \rightarrow Y$ (b) path does *not* represent a causal relation, even when three or more measurement occasions are used. It naturally follows, then, that the *ab*-mediated effect is also not a causal parameter. This is a basic problem of many intervention studies, described earlier in terms of violations of the sequential ignorability assumption. In this section, we discuss two potential design options for addressing this problem: enhancement designs and blockage designs. Some statistical methods to address this issue were mentioned previously (e.g., sensitivity analysis). Note that these methods are *design* approaches and focus on logical demonstration of mediation processes providing additional support for the presence or absence of a causal mechanism beyond what can be inferred from a traditional mediation model. We acknowledge that many mediators cannot be manipulated for practical or ethical reasons (e.g., family dysfunction); in these cases, the relation of M to Y must be based on existing conceptual theory for how M affects Y.

A **blockage design** uses an experimental manipulation to block the mediation process. If, as a result, the mediation relation is removed, then there is evidence for mediation. Blockage designs are most readily applied to studies examining a physiological process, such as reward pathways for addictive substances or brain chemical benefits of certain activities, but are often not possible to undertake for ethical reasons. Consider, for example, a study in which a treatment for depression that included an exercise component was administered to a full sample of people with major depressive disorder. If it were hypothesized that exercise reduces depression through increases in exercise-induced serotonin release, the researcher could "block" the release of serotonin in one group of subjects to determine whether serotonin release is a mediator. If serotonin release mediates the relation between exercise and depression, then depression would be higher in the group who received the serotonin release blocker.

A second approach to determining the presence of mediation with an experimental design is an **enhancement design**. In this case, exposure to the mediator is directly manipulated by increasing the dose of the mediator for one or more groups. To take the converse of the blockage design example above, suppose that depression was hypothesized to reduce physical activity, which, in turn, is positively related to immune functioning (see Miller, Cohen, & Herbert, 1999). To experimentally manipulate the mediator, one could obtain a sample of individuals who were equally depressed and assign them to moderate and heavy levels of physical activity. If mediation exists, then immune functioning should be higher for groups who were assigned to higher levels of physical activity.

The use of experimental design for investigating mediation processes is not new (Mark, 1986), but these issues have received more attention recently (MacKinnon, 2008; MacKinnon & Pirlott, 2009; Spencer, Zanna, & Fong, 2005). Experimental design approaches to treatment mediation research seems especially pertinent given that there are often a number of components within a particular treatment that can potentially be tested. However, there are often countervailing cost and practical issues involved in testing many components of a treatment.

Exemplar Clinical Mediation and Moderation Study

The investigation of mediation and moderation requires a range of research designs and

corresponding information. Nevertheless, there are several ideal characteristics of research design for investigating these variables. The quality of a treatment research project can be improved by decisions made before the study is conducted, as described in MacKinnon (2008). The ideal mediation study is based on detailed action and conceptual theory for how the treatment will achieve its effects. It is useful to construct an action theory table that relates treatment actions to mediators, ideally with some measure of effect size in the cells of the table, and also a conceptual theory table that includes measures of effect size for the relation of the mediators to the outcome variable. In the many cases where prior empirical research and theory relating mediators and moderators to outcomes is incomplete, the researcher must decide upon the ideal mediators and moderators. In an ideal research study, different theories for the process by which treatment affects outcomes are carefully specified such that they can be compared using the same dataset. Similarly, predictions regarding how processes will differ for different groups of clients are specified prior to the study.

The mediation model implies a temporal sequence among variables. As a result, hypotheses of how variables in the study will change over time, how variables will be related over time, and when an intervention will affect variables in the study are important to specify before the study begins, both for planning data analysis and also so that variables are measured at the right times to detect effects. Longitudinal measurement is important in treatment research for assessing change over time and for assessing the extent to which change in one variable predicts change in another—for example, if treatment-induced change in negative attributions leads to reduced depression. Similarly, it would be ideal to obtain measures at several points before and after treatment to estimate the baseline rate of change and the duration of effects. So an ideal treatment study would include at least two measurements before treatment is delivered and four measures after a study. Potential moderator variables should be specified and plans enacted to measure them; variables to consider include characteristics not likely to change, such as age, gender, and location, as well as baseline measures of characteristics that can change, such as the tendency to take risks, depression, and educational achievement. The extent to which baseline measures of outcome and mediator variables moderate treatment effects should also be considered; these measures will be available when a longitudinal design is used.

The validity and reliability of measures is critical for mediation and moderation analysis. The consequences of poor measurement include low power to detect effects, spurious relations among variables owing to shared method bias, and not measuring what is expected. One decision is whether to use narrow measures of a construct or more general measures, for example general self-efficacy or self-efficacy to avoid negative cognitive attributions. Qualitative measures of mediating and moderating processes may also be useful to obtain from at least some study participants. Obviously, it is important to obtain measures of mediators that are most likely to reflect the mechanism of change in treatment.

Several statistical issues are critical for investigation of mediating processes. In particular, sample size should be sufficient to be able to detect a real effect if it is there. With small sample sizes, real effects are unlikely to be detected. Generally, without baseline measurement, a sample size of 539, 74, and 35 is necessary to detect small, medium, and large effects (for both the *X*-to-*M* and *M*-to-*Y*, adjusted for *X*, paths) of a treatment (Fritz & MacKinnon, 2007). Large effects will require correspondingly fewer subjects. If a more complicated mediation model is hypothesized, it would be useful to assess power with a Monte Carlo method. Thoemmes, MacKinnon, and Reiser (2010) describe a Monte Carlo method to assess power for complex designs.

Random assignment to treatment conditions is important because it improves causal interpretation of some study results. It would also be useful to consider alternative designs to improve causal interpretation of a research study. Two types of designs were described in this chapter, the enhancement and blockage designs, although other designs are possible. In these designs the targeted mediator is enhanced or blocked in an additional treatment condition. A pattern of results across the treatment conditions would then provide evidence for or against the mediating process as critical for successful treatment. There are additional designs that may be useful here (Mackinnon & Pirlott, 2009). It is also important to include comparison mediators that would serve to improve interpretations of a mediation result. Comparison mediators are variables that are not expected to mediate the treatment effect but should be affected by the same confounding variables and error that influence the mediator measuring the actual mechanism. In this way, evidence for mediation through the true mediator should be observed, but not for the comparison mediator.

A final important aspect of the ideal study would be the detailed reporting of the specific model tested, along with confidence intervals for the mediation quantities, *a*, *b*, *c'*, and *ab*. It would be useful to employ the same best methods to calculate these (MacKinnon, 2008) across different studies. In this way, studies can be compared in the same way on variables with the same metric. Similarly, a detailed description of the measures used is critical. If this information cannot be included in publications, an Internet report could be prepared for all treatment studies, making it easier to combine information and obtain meaningful consistencies across studies. A look at existing treatment mechanism research shows that a wide variety of methods is used, many of which are not optimal; this makes it difficult to ascertain whether evidence for a mediating mechanism was observed.

In summary, the ideal mediation and moderation study has clearly specified mediational processes based on action and conceptual theory and clear specification of moderating variables prior to conducting the study. The study would include random assignment of clients to conditions in a longitudinal design, with the timing of observed measurements based on theory. Ideally at least four measurement points would be available to apply modern longitudinal methods and assess trends over time; two or more measurements before treatment would add to the interpretability of research results. Ideally the validity and reliability of each measure would be based on extensive empirical research. Additional research conditions that correspond to manipulating the potential mediating variables would also add credibility to the mediation hypothesis. The study would also be improved with the selection of comparison mediators that are not targeted by the intervention but are likely to be affected by the same confounders as the targeted mediator. Finally, mediation effects replicated and extended in future studies are necessary, as well as additional information relevant to identifying true mediating processes, such as qualitative data.

New Developments

New developments in the application of statistical mediation analysis continue at a rapid pace. Some of the most recently developed methods include complex hierarchical mediation models for dyadic data (Dagne, Brown, & Howe, 2007), Bayesian approaches (Yuan & MacKinnon, 2009), models for adaptive interventions (Collins et al., 2011), and meta-analytic approaches. We focus on Bayesian approaches, models for adaptive interventions, and meta-analytic approaches because they are particularly relevant for treatment research.

Bayesian Mediation Analysis

Thus far, this chapter has only outlined statistical methods within the frequentist paradigm of statistics. The Bayesian paradigm is another framework for statistics that is becoming more widely applied. One of the strengths of the Bayesian paradigm is the ability to incorporate prior information from previous studies in the current analysis. This feature is particularly attractive for researchers who wish to test treatments for rare disorders because such studies often rely on small sample sizes. To perform a Bayesian mediation analysis using prior information, the slopes and variances of the *a* and *b* paths from earlier studies can be applied to a model that uses the *Bayesian credibility interval* to allow for a skewed distribution, which is often a characteristic of small sample sizes (Pirlott, Kisbu-Sakarya, DeFrancesco, Elliot, & MacKinnon, 2012). Moreover, the credibility interval more closely reflects the nonnormal distribution of the product of *a* and *b*. Bayesian mediation analysis is also straightforward in complex mediation models (Yuan & MacKinnon, 2009) and is available in the Mplus (Muthén & Muthén, 2010) and WinBUGS software packages. However, to date there have been few applications of Bayesian methods for mediation analysis, so it would be ideal if more researchers would apply this method.

Models for Adaptive Interventions

Because the return to a healthy state in response to a treatment is typically not a discrete event, health psychology theorists have advocated for viewing recovery from a disorder or addiction as a process with several stages (Baker et al., 2010). A prominent example of this approach is Prochaska and colleagues' Transtheoretical Model of Behavior Change (Prochaska, Wright, & Velicer, 2008), which is rooted in the idea that the best treatments are adjusted according to what the client needs, when he or she needs it.

Variations of this approach have been widely applied, for example, in the smoking cessation literature. Recently, Baker and colleagues (2010) outlined four stages of cessation: motivation, precessation, cessation, and maintenance. Each of these stages has specific (yet often overlapping) needs, and clinical goals are adjusted according to these needs within this framework. For example, individuals in the motivation stage likely need treatment

programming components that increase their self-efficacy to quit smoking, which may become less important as the client progresses through the stages of change. Because treatment goals are different at the various stages, it follows that the mediators will also change (Baker et al., 2010) and should be adjusted accordingly.

The dynamic nature of phase-based treatment research is a challenge to estimate statistically, although advances in statistical modeling packages such as Mplus (Muthén & Muthén, 2010) have made this increasingly possible. The LCM described earlier is useful for adaptive interventions because it allows modeling of change at different times. Other approaches to adaptive modeling focus on the use of covariates and other information to obtain estimates of effects even though treatments change through the course of the study (Murphy, Van der Laan, & Robins, 2001).

Meta-analytic Approaches

One way to investigate mediational process from multiple treatment studies is to conduct a systematic review of studies relevant to action and conceptual theory in a research area and also quantify effect sizes for relations for the *X*-to-*M* and *M*-to-*Y* relations (MacKinnon, 2008). A methodology for combining quantitative information from multiple treatment studies is called *meta-analysis* (see also Chapter 17 in this volume) or integrated data analysis, which involves combining information from many treatment studies. There are several examples of this type of meta-analysis, including studies of the effects of psychotherapy (Shadish, 1996; Shadish & Baldwin, 2003; Shadish, Matt, Navarro, & Phillips, 2000; Shadish & Sweeney, 1991). A goal of mediation meta-analysis is to combine information across studies on the relations in a mediation model, *X* to *M* and *M* to *Y*, so as to obtain the most accurate estimates of these relations by combining information across many studies. In the context of treatment studies, the purpose of mediational meta-analysis is to determine the extent to which studies support conceptual and action theory for the intervention. Studying two relations, *X* to *M* and *M* to *Y*, in mediation meta-analysis is considerably more complicated than meta-analysis of a single *X*-to-*Y* relation, as in most meta-analyses. In addition, assuming *X* is randomized, there is a much stronger basis for the causal interpretation of the *X*-to-*M* relation than the *M*-to-*Y* relation, which is not randomized. Another complication in mediation meta-analysis is that information on just *X* to *M* or just *M* to *Y* or both *X* to *M* and *M* to *Y* may be available from a research study. Other limitations of mediation meta-analysis compared to regular meta-analysis include different measures across studies and different samples of participants, if the goal is to compare results across several treatments. Nevertheless, meta-analysis of mediation treatment studies is an ideal way to combine information across studies to improve treatments and the delivery of treatments to the best groups of individuals. For mediation and moderation meta-analyses, it is critical that individual studies report relations among variables that could be used for subsequent mediational analysis, such as values of *a*, *b*, *c*, and *c'* along with their standard errors and estimates of moderator effects and standard errors. Thorough descriptions of samples and measures used would further improve meta-analysis of mediating and moderating processes.

Summary and Future Research

Mediating mechanisms are fundamental to clinical treatment research for several reasons. A critical aspect of attention to mediating mechanisms is its focus on theory. Indeed, the purpose of mediating variables analysis is to directly test theory by explicitly obtaining estimates of quantities representing the extent to which a treatment worked by operating through a mediating process. Action theory, the relation of treatment actions to the mediator, and conceptual theory, the relation of the mediator to the outcome, are two important theoretical components of the mediation approach in treatment research. Attention to action and conceptual theory in treatment research is important in the theoretical development of treatments and their evaluation and may suggest that certain mediators are unlikely to be changeable given the actions available in treatment. However, treatments based on hunches without claims of mechanism of change may be successful, and then subsequent research may focus on explaining how the treatment achieved its effects. In any case, theory is a crucial aspect of the analysis of mediating variables.

Moderating variables are also fundamental to treatment research because treatments are developed for subgroups of persons with a certain disorder. Still, even when the treatment is designed for a certain subgroup of persons, there may well be certain client characteristics or contexts that interact with treatment, such that treatments have differential effects. The investigation of these moderating variables is an important aspect of clinical treatment design,

requiring both theoretical predictions and design of studies to be able to detect moderator effects if they exist. In particular, moderator effects may require larger sample sizes to detect effects because the power to detect interaction effects is reduced compared to the power to detect main effects.

New, more accurate approaches to statistical mediation analysis have been developed over the past several decades, and this work continues. Regarding statistical tests of mediating variables, the most accurate methods accommodate the nonnormal distribution of the estimate of the mediated effect, including resampling methods and the distribution of the product (MacKinnon, Lockwood, & Williams, 2004). A variety of longitudinal models are available for assessing mediation effects in treatment research that allow for individual differences in change as well as modeling group processes of change, such as the latent change model and the latent growth model. The mediation model is by definition a longitudinal model in which *X* causes *M* and *M* causes *Y*. More defensible decisions about temporal precedence can be made with data that are collected longitudinally.

There is a fundamental interpretation problem for the case of mediation analysis even when *X* represents random assignment to conditions. When *X* represents random assignment to conditions, the relations of *X* to *M* and *X* to *Y* can be interpreted as causal effects because randomization balances confounders of these relations. Randomization does not balance confounders of the *M*-to-*Y* relation, however, so causal interpretation of that mediator as the true mediator is more complicated. There are several ways that a researcher can address this limitation. One is to plan a program of research that sequentially tests more refined aspects of the mediation hypothesis, including, potentially, randomized experiments that more directly manipulate the mediator. The use of comparison mediators is also useful to show mediation for one mediator but not another. Generally, the accumulation of evidence from experimental data, qualitative data, and a variety of other sources builds the case for an important mediating mechanism. Causal inference for mediation relations is an active area of research likely to generate more useful methods over the next decade.

In summary, mediation analysis addresses the fundamental aspect of theories that explains processes by which treatments affect an outcome variable, and moderation analysis clarifies different effects of treatments across groups. Mediation and moderation analyses have the potential to answer theoretical questions commonly posed in treatment research, thereby reducing costs of treatments and increasing the scientific understanding of how treatments affect behavior. This chapter provided an overview of methods and issues involved in identifying mediating and moderating variables. The ideal next step is the repeated application of these methods with real data.

Note

This article was supported in part by Public Health Service grant DA09757 from the National Institute on Drug Abuse. We thank George Howe and Linda Luecken for comments on the chapter and MH40859 from the National Institute on Mental Health.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*, 69–85.

Baker, T. B., Mermelstein, R., Collins, L. M., Piper, M. E., Jorenby, D. E., Smith, S. S., Christiansen, B. A., Schlam, T. R., Cook, J. W., & Fiore, M. C. (2011). New methods for tobacco dependence treatment research. *Annals of Behavioral Medicine*, *41*, 192–207.

Barber, J. P., & DeRubeis, R. J. (1989). On second thought: Where the action is in cognitive therapy for depression. *Cognitive Therapy and Research*, *13*, 441–457.

Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the Treatment for Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology*, *64*, 951–958.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Bentler, P. M. (1997). *EQS for Windows (version 5.6)* [Computer program]. Encino, CA: Multivariate Software, Inc.

Bricker, J. B., Liu, J., Comstock, B. A., Peterson, A. V., Kealey, K. A., & Marek, P. M. (2010). Social cognitive mediators of adolescent smoking cessation: Results from a large randomized intervention trial. *Psychology of Addictive Behaviors*, *24*, 436–445.

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Cheong, J., MacKinnon, D. P., & Khoo, S. T. (2003). Investigation of mediational processes using parallel process latent growth curve modeling. *Structural Equation Modeling*, *10*, 238–262.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Routledge Academic.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558–577.

Collins, L. M., Baker, T. B., Mermelstein, R. J., Piper, M. E., Jorenby, D. E., Smith, S. S., Christiansen, B. A., et al. (2011). The multiphase optimization strategy for engineering effective tobacco use interventions. *Annals of Behavioral Medicine*, *41*, 208–226.

Collins, L. M., Graham, J. W., & Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, *33*, 295–312.

Dagne, G. A., Brown, C. H., & Howe, G. W. (2007). Hierarchical modeling of sequential behavioral data: Examining complex association patterns in mediation models. *Psychological Methods*, *12*, 298–316.

DeRubeis, R. J., Evans, M. D., Hollon, S. D., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1990). How does cognitive therapy work? Cognitive change and symptom change in cognitive therapy and pharmacotherapy for depression. *Journal of Consulting and Clinical Psychology*, *58*, 862–869.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive therapy vs. medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, *62*, 409–416.

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., et al. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*, *74*, 658–670.

Doss, B. D. (2004). Changing the way we study change in psychotherapy. *Clinical Psychology: Science and Practice*, *11*, 368–386.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1–22.

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH Treatment of Depression Collaborative Research Program: Background and research plan. *Archives of General Psychiatry*, *42*, 305–316.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, *10*, 87–99.

Ferrer, E., & McArdle, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling*, *10*, 493–524.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Gallop, R., Amsterdam, J. D., & Hollon, S. D. (2008). Antidepressant medications v. cognitive therapy in people with depression with or without personality disorder. *British Journal of Psychiatry*, *192*, 124–129.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Freedheim, D. K., & Russ, S. W. (1992). Psychotherapy with children. In C. Walker & M. Roberts (Eds.), *Handbook of clinical child psychology* (2nd ed., pp. 765–781). New York: Wiley.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.

Gaynor, S. T., & Harris, A. (2008). Single-participant assessment of treatment mediators: Strategy description and examples from a behavioral activation intervention for depressed adolescents. *Behavior Modification*, *32*, 372–402.

Gibbons, M. B. C., Crits-Christoph, P., Barber, J. P., Wiltsey Stirman, S., Gallop, R., Goldstein, L. A., et al. (2009). Unique and common mechanisms of change across cognitive and dynamic psychotherapies. *Journal of Consulting and Clinical Psychology*, *77*, 801–813.

Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 243–259). Washington DC: American Psychological Association.

Happé, F., & Ronald, A. (2008). The "fractionable autism triad": a review of evidence from behavioural, genetic, cognitive and neural research. *Neuropsychology Review*, *18*, 287–304.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.

Hyman, H. H. (1955). *Survey design and analysis: Principles, cases, and procedures*. Glencoe, IL: Free Press.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309–334.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*, 51–71.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL (Version 8.5)* [Computer program]. Chicago, IL: Scientific Software International, Inc.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.

Kaskutas, L. A., Bond, J., & Humphreys, K. (2002). Social networks as mediators of the effect of Alcoholics Anonymous. *Addiction (Abingdon, England)*, *97*, 891–900.

Kazdin, A. E. (1989). Childhood depression. In E. J. Mash & R. A. Barkley (Eds.) *Treatment of childhood disorders* (pp. 135–166). New York: Guilford Press.

Kazdin, A. E. (2000). Developing a research agenda for child and adolescent psychotherapy. *Archives of General Psychiatry*, *57*, 829–835.

Kazdin, A. E. (2002). *Research design in clinical psychology* (4th ed.). Needham Heights, MA: Allyn & Bacon.

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.

Kelly, J.F., Stout, R. L., Magill, M., Tonigan, J. S., & Pagano, M. E., (2011). Spirituality in recovery: A lagged mediational analysis of Alcoholics Anonymous' principal theoretical mechanism of behavior change. *Alcoholism Clinical and Experimental Research*, *35*, 454–463.

Kendall, P. C., & Treadwell, K. (2007). The role of self-statements as a mediator in treatment for anxiety-disordered youth. *Journal of Consulting and Clinical Psychology*, *75*, 380–389.

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes. *Journal of the American Medical Association*, *296*, 1286–1289.

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*, S101–S108.

Kraemer, H. C., Wilson, T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–883.

Lockhart, G. L., MacKinnon, D. P., & Ohlirch, V. (2011). Mediation analysis in psychosomatic medicine research. *Psychosomatic Medicine*, *73*, 29–43.

Longabaugh, R., & Magill, M. (2011). Recent advances in behavioral addiction treatments: Focusing on mechanisms of change. *Current Psychiatry Reports*, *13*, 382–389.

Longabaugh, R., & Writz, P. W. (Eds.) (2001). *Project MATCH hypotheses: Results and causal chain analyses*. U.S. Department of Health and Human Services. Washington, DC: U. S. Government Printing Office.

Luborsky, L., Mark, D., Hole, A. V., Popp, C., Goldsmith, B., & Cacciola, J. (1995). *Supportive-expressive dynamic psychotherapy of depression: A time-limited version.* New York: Basic Books.

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. In A. Cazares & L. A. Beatty (Eds.), *Scientific methods for prevention/intervention research* (NIDA Research Monograph Series 139, DHHS Pub 94–3631, pp. 127–153). Washington, DC: U. S. Department of Health and Human Services.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

MacKinnon, D. P. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice* doi:10.1177/1049731511414148

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144–158.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*, 384–389.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, *1*, 173–181.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). Comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128.

MacKinnon, D. P., & Luecken, L. J. (2011). Statistical analysis for identifying mediating variables in public health dentistry interventions. *Journal of Public Health Dentistry*, *71*, S37–S46.

MacKinnon, D. P., & Pirlott, A. (2009). *The unbearable lightness of b*. Paper presented at the Annual Meeting of the Society for Personality and Social Psychology.

MacKinnon, D. P., Weber, M. D., & Pentz, M. A. (1989). How do school-based drug prevention programs work and for whom? *Drugs and Society*, *3*, 125–143.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 47–66). San Francisco: Jossey-Bass.

McCrady, B. S., Epstein, E. E., Cook, S., Jensen, N., & Hildebrandt, T. (2009). A randomized trial of individual and couple behavioral alcohol treatment for women. *Journal of Consulting and Clinical Psychology*, *77*, 243–256.

Merrill, R. M. (1994). *Treatment effect evaluation in non-additive mediation models*. Unpublished dissertation, Arizona State University.

Miller, G. E., Cohen, S., & Herbert, T. B. (1999). Pathways linking major depression and immunity in ambulatory female patients. *Psychosomatic Medicine*, *61*, 850–860.

Moyers, T. B., Martin, T., Houck, J.M., Christopher, P. J., & Tonigan, J. S. (2009). From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of Consulting and Clinical Psychology*, *77*, 1113–1124.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and when mediation is moderated. *Journal of Personality and Social Psychology*, *89*, 852–863.

Murphy, S. A., Van der Laan, M. J., Robins, J. M., & Conduct Problems Prevention Research Group. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, *96*, 1410–1423.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Nock, M. K. (2007). Conceptual and design essentials for evaluating mechanisms of change. *Alcoholism Clinical and Experimental Research, 31*, 4S–12S.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*, 109–118.

Pini, S., Cassano, G. B., Simonini, E., Savino, M., Russo, A., & Montgomery, S. A. (1997). Prevalence of anxiety disorders comorbidity in bipolar depression, unipolar depression and dysthymia. *Journal of Affective Disorders, 42*, 145–153.

Piper, M. E., Federmen, E. B., McCarthy, D. E., Bolt, D. M., Smith, S. S., Fiore, M. C., & Baker, T. B. (2008). Using mediational models to explore the nature of tobacco motivation and tobacco treatment effects. *Journal of Abnormal Psychology*, *117*, 94–105.

Pirlott, A. G., Kisbu-Sakarya, Y., DeFrancesco, C. A., Elliot, D. L. & MacKinnon, D. P. (2012) *Mechanisms of motivational interviewing in health promotion: a Bayesian mediation analysis. International Journal of Behavioral Nutrition and Physical Activity*, *9*, 69.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating theories of health behavior change: A hierarchy of criteria applied to the transtheoretical model. *Applied Psychology*, *57*, 561–588.

Ratajczak, H. V. (2011) Theoretical aspects of autism: Causes—A review. *Journal of Immunotoxicology*, *8*, 68–79.

Resick, P. A., Galovski, T. E., O'Brien Uhlmansiek, M., 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*, 243–258.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, *1*, 47–65.

Shadish, W. R., & Baldwin, S. A. (2003). Meta-analysis of MFT interventions. *Journal of Marital and Family Therapy*, *29*, 547–570.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*, 512–529.

Shadish, W. R., & Sweeney, R. B. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, *59*, 883–893.

Shiffman S. (1984). Coping with temptations to smoke. *Journal of Consulting and Clinical Psychology*, *52*, 261–267.

Sotsky, S. M., Glass, D. R., Shea, M. T., & Pilkonis, P. A. (1991). Patient predictors of response to psychotherapy and pharmacotherapy: Findings in the NIMH Treatment of Depression Collaborative Research Program. *American Journal of Psychiatry*, *148*, 997–1008.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than meditational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845–851.

Stiles, W. B., & Shapiro, D. A. (1994). Disabuse of the drug metaphor: Psychotherapy process-outcome correlations. *Journal of Consulting and Clinical Psychology*, *62*, 942–948.

Tein, J. Y., Sandler, I. N., MacKinnon, D. P., & Wolchik, S. A. (2004). How did it work? Who did it work for? Mediation in the context of a moderated prevention effect for children of divorce. *Journal of Consulting and Clinical Psychology*, *72*, 617–624.

Ten Have, T. R., & Joffe, M. M. (2010). A review of causal estimation of effects in mediation analysis. *Statistical Methods in Medical Research*, published on-line.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analysis with rank preserving models. *Biometrics*, *63*, 926–934.

Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling*, *8*, 510–534.

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43*, 692–700.

Treadwell, K. R. H., & Kendall, P. C. (1996). Self-talk in anxiety-disordered youth: States-of-mind, content specificity, and treatment outcome. *Journal of Consulting and Clinical Psychology*, *64*, 941–950.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, *78*, 2957–2962.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, *21*, 540–551.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology*, *22*, 582–585.

von Eye, A., Mun, E. Y., & Mair, P. (2009). What carries a mediation process? Configural analysis of mediation. *Integration of Psychology and Behavior*, *43*, 228–247.

Weersing, V. R., & Weisz, J. R. (2002). Mechanisms of action in youth psychotherapy. *Journal of Child Psychology and Psychiatry*, *43*, 3–29.

Weisz, J. R., & Kazdin, A. E. (2003). Concluding thoughts: Present and future of evidence-based psychotherapies for children and adolescents. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 439–452). New York: Guilford.

Williams, G. C., McGregor, H. A., Sharp, D., Levesque, C., Kouides, R. W., Ryan, R. M., & Deci, E. L. (2006). Testing a self-determination theory intervention for motivating tobacco cessation: supporting autonomy and competence in a clinical trial. *Health Psychology*, *25*, 91–101.

Yuan, Y., & Mackinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322.

CHAPTER

16 Structural Equation Modeling: Applications in the Study of Psychopathology

Erika J. Wolf *and* Timothy A. Brown

Abstract

Structural equation modeling (SEM) has become increasingly popular among social science researchers, yet many applied clinical researchers are hesitant to utilize this powerful and flexible multivariate data analytic technique. This chapter provides an introductory and applied guide to the use of common SEM models, with a focus on how SEM can help advance the understanding of psychopathology and its treatment. The chapter introduces general SEM concepts such as model depiction, identification, and evaluation; it also describes the advantages associated with this approach to data analysis. The chapter presents specific, applied examples of confirmatory factor analysis, structural regression models, and latent growth models. Sample Mplus scripts and output are provided.

Key Words: Multivariate analysis, structural equation modeling, latent variable, regression analysis, confirmatory factor analysis, latent growth model

Structural equation modeling (SEM) has become increasingly popular among social science researchers. This likely reflects greater awareness of the power of this type of multivariate data analysis to inform our conceptualization of psychological constructs as well as greater accessibility of statistical modeling software programs. Despite this, SEM maintains a mystique of being too complicated and unattainable for many applied clinical researchers. We believe this notion is unwarranted, and our aim is to provide an introductory and applied guide to the use of common SEM models and to discuss how SEM can help advance the understanding of psychopathology and its treatment. We begin the chapter by providing a basic orientation to SEM, including discussion of model depiction, identification, and evaluation. We then discuss the utility of specific types of SEM, including *confirmatory factor analysis* (CFA), *structural regression models*, and *latent growth curve models*. We provide applied examples of these approaches and some sample syntax and discuss the potential role of each approach in advancing psychopathology research and psychological science. Those interested in more technical aspects of SEM, please see Brown (2006) and Kline (2010).

What Is SEM?

SEM is an umbrella term that refers to a type of path analysis in which one or more constructs of interest is not directly measured, but is instead included in the model as a *latent variable*. A latent variable is defined by several observed variables (or *indicators*) that are highly correlated with one another and are presumed to measure the same construct. For example, common variance across self-report, interview, and behavioral measures of depression can be used together to define a latent variable reflecting the construct of depression. Through the use of multiple indicators of the latent construct of depression, researchers can distinguish true score variance (i.e., variance that is shared by the indicators) and error variance (i.e., unique

variance in the indicator that is not explained by the latent variable and that is typically presumed to be measurement error). At its heart, SEM is simply a form of regression whereby multiple regressions are conducted simultaneously. What makes SEM "structural" is its ability to analyze the interrelationships among latent variables in the context of the adjustment for measurement error and a measurement error theory.

Why Use SEM?

There are several reasons why SEM is advantageous over single indicator-based analyses (e.g., ordinary least squares statistics such as multiple regression and analysis of variance). First, latent variable modeling allows for improved construct validity of the variables of interest. Let's return to our example of a latent variable of depression. The construct validity of a single indicator of depression is limited by the reliability of the measure, by the modality of the assessment (e.g., interview vs. self-report), and by its ability to adequately cover the content domain of the broader depression construct. It captures a slice from the depression pie. In contrast, a latent variable reflecting depression uses information from multiple slices of the depression pie, yielding improved coverage of the domain. In other words, the latent variable of depression is probably a closer representation of the construct of depression, as it exists "in nature," relative to any single measurement of depression. A second, related, point is that single indicators of depression, whether they be dimensional severity scores or diagnostic binary codes, contain both true score variance and error variance. The inclusion of error variance in the score leads to biased parameter estimates (i.e., regression coefficients) in a regression model. In contrast, latent variables are theoretically free of measurement error; they contain only true score variance because unique variance (i.e., variance in the indicators that is not explained by the latent variables) has been removed and is modeled as indicator error. This yields more accurate parameter estimates of the associations among constructs in the model.

Third, the ability to separate true score from error variance affords a number of options for researchers, including the ability to model correlated error in instances in which the errors from two or more indicators might correlate with one another for methodological reasons (e.g., a method effect such as the same assessment modality or similar item wording). Fourth, the improved reliability and construct validity of a latent variable means that the model, as a whole, has greater statistical power to detect the effects of interest. Fifth, SEM allows for the inclusion of multiple independent and dependent variables in the same analysis and simultaneously solves for all equations in the model. For example, in a mediation analysis, rather than conducting separate regression equations to determine the effect of a mediator on the association between an independent and dependent variable, all direct and indirect paths in the SEM model are estimated at once. This allows for an elegant and parsimonious approach to the evaluation of complex models. Finally, confirmatory forms of SEM (e.g., CFA) allow researchers to test a hypothesized model against their data and, in instances where the researchers have alternative hypotheses, to directly compare the results of two or more models.

Model Representation

SEM lends itself well to graphical depictions of associations among variables. These images succinctly convey a great deal of information and are relatively easy to understand. Figures of SEM models follow a convention. As shown in the measurement model of Obsessions and Worry in Figure 16.1, indicators of a construct are depicted as squares and latent variables are shown as circles. Single-headed arrows going from the latent variable to the indicators are *factor loadings* (i.e., regression paths in which variability in the indicator is accounted for by the latent variable). Single-headed arrows pointing toward the other side of the indicators reflect *error variance*. This value is the variance in the indicator that is unaccounted for by the latent construct. The double-headed arrow in Figure 16.1 reflects the covariance of the two factors. As will be shown later, each of the parameter estimates in the model is interpreted in a manner analogous to regression.

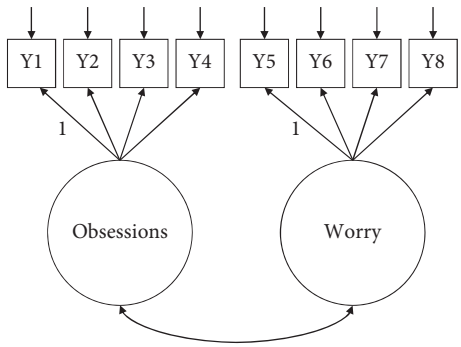
Model Identification

Before proceeding into a discussion of the use of SEM for studying the nature and treatments of psychopathology, it is important to understand the principles of *model identification*. An aim of SEM is to reproduce the associations among the data with the fewest number of freely estimated parameters so that parsimonious and efficient models are developed. To estimate an SEM solution, the model must be *identified*. A model is identified if, on the basis of known information (i.e., the variances and covariances in the sample input matrix), a unique set of estimates for each parameter in the model can be obtained (e.g., factor loadings, factor covariances, etc.). The two primary aspects of SEM model identification are scaling the latent variables and statistical identification.

Latent variables have no inherent metrics, and thus their units of measurement must be set by the researcher. In SEM, this is usually accomplished in one of two ways. The most widely used method is the *marker indicator* approach whereby the unstandardized factor loading of one observed measure per factor is fixed to a value of 1.0 (shown in Fig. 16.1). This specification passes the measurement scale of the marker indicator down to the latent variable. In the second method, the variance of the latent variable is fixed to a value of 1.0. Although most SEM results are identical to the marker indicator approach when the factor variance is fixed to 1.0 (e.g., goodness of fit of the solutions is identical), only completely and partially standardized solutions are produced. While perhaps useful in some circumstances, the absence of an unstandardized solution often contraindicates the use of this approach (e.g., in scenarios where the unstandardized regression paths among the latent variables have strong interpretative value).

Statistical identification refers to the concept that an SEM solution can be estimated only if the number of freely estimated parameters (e.g., factor loadings, error variances, factor covariances) does not exceed the number of pieces of information in the input matrix (e.g., number of sample variances and covariances). It is the sample variance–covariance matrix that forms the basis of the SEM analysis when the maximum likelihood (ML) estimator is used. So, for example, if we have variables A, B, and C in our dataset, we have six pieces of information in the input variance–covariance matrix: the variance of each variable ($n = 3$) and the covariance of each variable with the other variables ($n = 3$: A with B, A with C, and B with C). For models with more variables, a convenient way to calculate the number of pieces of information in a model is by using the following equation: $[p(p + 1)]/2$, where p is the number of indicators in the model. As long as we estimate fewer parameters relative to the number of pieces of information, the model is said to be *overidentified* and there is at least 1 degree of freedom (*df*) to use to solve the equations. With ML estimation, model *df* is the difference between the number of pieces of information in the model and the number of freely estimated parameters.

For example, the two-factor measurement model in Figure 16.1 is overidentified with $df = 19$. The model *df* indicates that there are 19 more elements in the input variance–covariance matrix than there are freely estimated parameters in this model. Specifically, there are 36 variances and covariances in the input matrix; $[8(8 + 1)]/2 = 36$. There are 17 freely estimated parameters in the model—that



Correlations/Standard Deviations (*SD*s):

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Y1	1.000							
Y2	0.529	1.000						
Y3	0.441	0.503	1.000					
Y4	0.495	0.658	0.509	1.000				
Y5	0.069	0.110	0.214	0.106	1.000			
Y6	0.141	0.140	0.169	0.152	0.425	1.000		
Y7	0.098	0.058	0.144	0.096	0.334	0.401	1.000	
Y8	0.097	0.138	0.138	0.126	0.252	0.219	0.186	1.000
*SD*s:	2.725	2.278	2.362	2.356	2.114	2.227	2.076	1.543

Figure 16.1 Confirmatory factor analysis model of Obsessions and Worry.

is, six factor loadings (the factor loadings of Y1 and Y5 are not included because they were fixed to 1.0 to serve as marker indicators), two factor variances, one factor covariance, and eight error variances (see Fig. 16.1). Thus, model $df = 19$ (36 − 17).

In some cases, the number of pieces of information in a model is equal to the number of freely estimated parameters in a model. This means that the $df = 0$, and the model is said to be *just-identified*. Although just-identified models can be estimated, goodness-of-fit evaluation does not apply because these solutions perfectly reproduce the input variance–covariance matrix. In just-identified models, there exists one unique set of parameter estimates that solve all the equations in the model. In other cases, the researcher may erroneously try to specify a model that contains more freely estimated parameters than available pieces of information in the input matrix. When the number of freely estimated parameters exceeds the number of pieces of information in the input matrix (e.g., when too many factors are specified for the number of indicators in the sample data), *df*s are negative and the model is *underidentified*. Underidentified models cannot be estimated because the solution cannot arrive at a unique set of parameter estimates. Latent variable software programs will output an error message if such a model is specified. In some cases, a model may be just-identified or overidentified from a statistical standpoint, but the model is said to be *empirically underidentified* due to how the model is specified, the pattern of associations in the data, or both. For example, a model with three indicators that load on a single latent variable would be just-identified; however, if two of the indicators have no relationship with the third indicator (i.e., $r = 0$), then the model is empirically underidentified and a solution cannot be obtained as there is insufficient information to generate a unique set of parameter estimates. The reader is referred to Brown (2006) and Wothke (1993) for more information on the causes and remedies for empirically underidentified solutions.

Model Evaluation

How do researchers determine the acceptability of their SEM model? The first step in evaluating a model is to carefully read through the output to determine that no serious errors occurred. A serious error would be indicated if the model fails to converge, if standard errors cannot be computed for the parameter estimates, or if some parameter estimates in the solution have out-of-range values (e.g., a negative residual variance). The model should not be evaluated further in the presence of any of these serious errors. Moreover, the researcher should carefully review the output to ensure the model was executed as intended. Both modeling program defaults and human error can result in unintended freely estimated parameters, or parameter estimates that were erroneously omitted or otherwise misspecified.

There are three major aspects of the results that should be examined to evaluate the acceptability of SEM models: (1) overall goodness of fit; (2) the presence or absence of localized areas of strain in the solution (i.e., specific points of ill fit); and (3) the interpretability, size, and statistical significance of the model's parameter estimates. Goodness of fit pertains to how well the parameter estimates of the solution (e.g., factor loadings, factor correlations) are able to reproduce the relationships observed in the sample data. There are a variety of goodness-of-fit statistics that provide a global descriptive summary of the ability of the model to reproduce the input variance–covariance matrix. The classic goodness-of-fit index is χ^2. Ideally, the model χ^2 would be statistically nonsignificant (i.e., does not exceed the critical value of the χ^2 distribution based on the *df* of the model), which would lead to the retention of the null hypothesis that the sample and model-implied variance–covariance matrices do not differ. Although χ^2 is steeped in the tradition of SEM (e.g., it was the first fit index to be developed), it is rarely used in applied research as a sole index of model fit. There are number of salient drawbacks of this statistic (e.g., see Brown, 2006), including the fact that it is highly sensitive to sample size (i.e., solutions involving large samples are often rejected on the basis of χ^2 even when differences between the sample and model-implied matrices are negligible). Nevertheless, χ^2 is used for other purposes such as nested model comparisons and the calculation of other goodness-of-fit indices. Thus, while χ^2 is routinely reported in SEM research, other fit indices are usually relied on more heavily in the evaluation of model fit.

In addition to χ^2, the most widely accepted global goodness-of-fit indices are the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), Tucker-Lewis index (TLI), and comparative fit index (CFI). The RMSEA reflects how well the model fits the data but does not require an exact match between the specified model and the data in the way that the χ^2 statistic does (i.e., it operates on a noncentral χ^2 distribution). It also takes into account model parsimony in that fit is estimated as a function of the *df* of the model. The SRMR

reflects a positive square root average of the discrepancy between the correlations among indicators that are implied by the model and the correlations that were observed in the sample data. The TLI and CFI both compare the fit of the specified model to a null model in which there are no associations among the variables of interest; in addition, the TLI invokes a penalty function when unnecessary parameters are included in the model. In one of the more comprehensive and widely cited evaluations of cutoff criteria, the findings of simulation studies conducted by Hu and Bentler (1999) suggest the following guidelines for acceptable model fit: (a) RMSEA values are close to .06 or below; (b); SRMR values are close to .08 or below; and (c) TLI and CFI values are close to .95 or greater. However, this topic continues to be strongly debated by methodologists (see Marsh, Hau, & Wen, 2004). For instance, some researchers assert that these guidelines are far too conservative for many types of models (e.g., measurement models composed of many indicators and several factors where the majority of cross-loadings and error covariances are fixed to zero; cf. Marsh et al., 2004). Moreover, because the performance of fit statistics and their associated cutoffs has been shown to vary as a function of various aspects of the model (e.g., degree of misspecification, size of factor loadings, number of factors; e.g., Beauducel & Wittman, 2005), the fit statistic thresholds suggested by simulation studies may have limited generalizability to many SEM models in applied research.

The second aspect of model evaluation is to identify whether there are specific areas of ill fit in the solution. A limitation of goodness-of-fit statistics (e.g., SRMR, RMSEA, CFI) is that they provide a *global*, descriptive indication of the ability of the model to reproduce the observed relationships among the indicators in the input matrix. However, in some instances, overall goodness of fit indices suggest acceptable fit despite the fact that some relationships among indicators in the sample data have not been reproduced adequately (or alternatively, some model-implied relationships may markedly exceed the associations seen in the data). This outcome is more apt to occur in complex models (e.g., models that entail an input matrix consisting of a large set of indicators) where the sample matrix is reproduced reasonably well on the whole, and the presence of a few poorly reproduced relationships has less impact on the global summary of model fit. On the other hand, overall goodness-of-fit indices may indicate that a model poorly reproduced the sample matrix. However, these indices do not provide information on the reasons why the model fit the data poorly (e.g., misspecification of indicator–factor relationships, failure to model salient error covariances).

Two statistics that are frequently used to identify specific areas of misfit in an SEM solution are *standardized residuals* and *modification indices*. A standardized residual is a standardized index (interpreted like a z score) of the difference between the model-implied association between any two variables in the analysis and the observed association between the same two variables. This statistic allows researchers to evaluate whether the relationships between variables are underestimated or overestimated by the model (as evidenced by positively and negatively signed standardized residuals, respectively). Stated another way, these values can be conceptually considered as the number of standard deviations that the residuals differ from the zero-value residuals that would emanate from a perfectly fitting model. For instance, a standardized residual at a value of 1.96 or higher would show there exists significant additional covariance between a pair of indicators that was not reproduced by the model's parameter estimates (i.e., 1.96 is the critical value of the z distribution, $\alpha = .05$, two-sided).

Modification indices can be computed for each fixed parameter (e.g., parameters that are fixed to zero such as indicator cross-loadings; cf. Fig. 16.1) and each constrained parameter in the model (e.g., parameter estimates that are constrained to be same the value). The modification index reflects an approximation of how much the overall model χ^2 will decrease if the fixed or constrained parameter is freely estimated. Because the modification index can be conceptualized as a χ^2 statistic with 1 *df*, indices of 3.84 or greater (i.e., the critical value of χ^2 at $p < .05$, 1 *df*) suggest that the overall fit of the model could be significantly improved if the fixed or constrained parameter was freely estimated. Because modification indices are sensitive to sample size, software programs provide expected parameter change (EPC) values for each modification index. As the name implies, EPC values are an estimate of how much the parameter is expected to change in a positive or negative direction if it were freely estimated in a subsequent analysis. Although standardized residuals and modification indices provide specific information for how the fit of the model can be improved, such revisions should be pursued only if they can be justified on empirical or conceptual grounds (e.g., MacCallum, Roznowski, & Necowitz, 1992). Atheoretical specification searches

(i.e., revising the model solely on the basis of large standardized residuals or modification indices) will often result in further model misspecification and overfitting (e.g., inclusion of unnecessary parameter estimates due to chance associations in the sample data; MacCallum, 1986).

The final major aspect of SEM model evaluation pertains to the interpretability, strength, and statistical significance of the parameter estimates. The parameter estimates (e.g., factor loadings and factor correlations) should be interpreted only in context of a good-fitting solution. If the model did not provide a good fit to the data, the parameter estimates are likely biased (incorrect). In context of a good-fitting model, the parameter estimates should first be evaluated to ensure they make statistical and substantive sense. As discussed earlier, the parameter estimates should not take on out-of-range values (often referred to as *Heywood cases*) such as a negative indicator error variance. These results may be indicative of a model specification error or problems with the sample or model-implied matrices (e.g., a nonpositive definite matrix, small *N*). Thus, the model and sample data must be viewed with caution to rule out more serious causes of these outcomes (see Wothke, 1993, and Brown, 2006, for further discussion). From a substantive standpoint, the parameters should be of a magnitude and direction that is in accord with conceptual or empirical reasoning. Small or statistically nonsignificant estimates may be indicative of unnecessary parameters. On the other hand, extremely large parameter estimates may be substantively problematic. For example, if in a multifactorial solution the factor correlations approach 1.0, there is strong evidence to question whether the latent variables represent distinct constructs (i.e., they have poor discriminant validity). As noted in the discussion of standardized residuals and modification indices, the model can be revised (e.g., removal of a nonsignificant path) provided that the revisions are justified on empirical or conceptual grounds. Otherwise, the researcher runs the risk of embarking on a post hoc specification search whereby the model specification is driven primarily by the sample data.

Programs to Evaluate SEMs

We will next discuss types of SEM and their application to the study of psychopathology and its treatment. There are a number of programs available to evaluate SEMs. These include Mplus, LISREL, EQS, Amos, and R. Because of its current popularity, the ensuing sections will include examples of Mplus syntax and output for common types of SEM models. Like other latent variable software programs, Mplus can handle many different types of data for use in SEM (e.g., dimensional indicators, dimensional indicators that are not normally distributed, count variables, dichotomous variables, ordered polytomous variables, censored data, data with missing values). The nature of the data will determine the most appropriate type of estimator to use for the analysis. For example, maximum likelihood (ML) is the most commonly used estimator but assumes the continuous data are multivariate normal. When the sample data violate the assumption of normality, other estimators should be used to avoid serious bias in the goodness-of-fit statistics, standard errors, and even the parameter estimates themselves (e.g., MLR for nonnormal continuous data, WLSMV for categorical outcomes). Many of the leading statistical estimators (e.g., ML, MLR) can accommodate missing data in a manner that is superior to traditional methods (e.g., listwise or pairwise deletion) and that does not require additional data processing outside of the latent variable analyses (e.g., multiple imputation; cf. Allison, 2003).

CFA: Application to the Measurement of Psychopathology

CFA has many applications in psychopathology research. This approach can be used to validate measures and constructs, to demonstrate the discriminant validity of constructs, and to generally improve the quality of the measurement of psychological constructs. CFAs are referred to as "measurement models" because CFA is primarily concerned with the relationships between observed measures (indicators) and their underlying constructs (e.g., as depicted in Fig. 16.1, directional paths are not specified among the latent variables). As noted earlier, because the latent variable is a better approximation of the true construct of interest than individual observed measures, it is associated with improved reliability, validity, and statistical power. In turn, this increases the researcher's likelihood of obtaining significant correlates, such as biomarkers for diseases, environmental risk factors, or treatment effects of psychopathological constructs. At a broader level, this can help inform our understanding of the etiology and course of psychopathology, as well as our conceptual nosology of psychological disorders. In this section we will review some examples of how CFA can be used for such purposes.

CFA Model with Two Correlated Factors

We begin with a basic example of CFA: a measurement model with two latent variables and eight

indicators, as shown in Figure 16.1. In this simulated example, eight measures (0–8 rating scales) of the symptoms of obsessive thinking and excessive worry were obtained from 400 outpatients with anxiety and mood disorders. Although we use a fairly large sample in this example, readers should not take this to mean that CFA (or other SEM analyses) can be evaluated only in very large samples. There is no "one size fits all" rule of thumb for sample size requirements for CFA (or SEM) because the minimum sample size depends on the particular aspects of the model and dataset, such as the strength of the relationships among measures (e.g., magnitude of factor loadings and correlations), model complexity (e.g., number of indicators and factors), type of data (e.g., continuous or categorical outcomes), research design (e.g., cross-sectional or longitudinal data), and extent of data nonnormality and missingness. Under favorable conditions (e.g., large effect sizes, multivariate normal outcomes), simulation research has indicated that many cross-sectional (e.g., CFA) and time-series (e.g., latent growth models) SEM models perform well with sample sizes well under 100. However, much larger samples are needed under other conditions (e.g., categorical outcomes, weaker relationships among indicators). Readers are encouraged to see Brown (2006) and Muthén and Muthén (2002) for details on conducting Monte Carlo simulation studies to determine the necessary sample size to ensure adequate statistical power and precision of the parameter estimates for a specific model of interest.

In the simulated example, a two-factor model is anticipated; that is, the first four measures (Y1–Y4) are conceptualized as indicators of the latent construct of Obsessions, and the remaining four indicators (Y5–Y8) are conjectured to be features of the underlying dimension of Worry. Although the constructs of Obsessions and Worry are predicted to be distinct, they are nonetheless expected to be correlated (as reflected by the double-headed arrow). In this model, none of the indicator error variances are correlated with one another, although error covariances can be specified if there is reason to believe the indicators are correlated with one another for reasons other than the latent variables (e.g., method effects stemming from measurement artifacts such as indicators with overlapping item content).

Figure 16.1 also provides the sample data for the two-factor measurement model—that is, the sample standard deviations (*SD*) and correlations (*r*) for the eight indicators. These data will be read into Mplus and converted into variances and covariances, which will be used as the input matrix (e.g., $VAR_{X1} = SD_{X1}^2$; $COV_{X1,X2} = r_{X1,X2}SD_{X1}SD_{X2}$). Generally, it is preferable to use a raw data file as input for SEM (e.g., to avoid rounding error and to adjust for missing or nonnormal data, if needed). In this example, however, the sample *SD*s and *r*s are presented so they can be readily used as input if the reader is interested in replicating the analyses presented in this chapter.

The Mplus syntax and selected output for the two-factor measurement model are presented in Table 16.1. As shown in the Mplus syntax, both the indicator correlation matrix and standard deviations (TYPE = STD CORRELATION;) were input because this CFA analyzed a variance–covariance matrix. The CFA model specification occurs under the "MODEL:" portion of the Mplus syntax. For instance, the line "OBS BY Y1-Y4" specifies that a latent variable to be named "OBS" (Obsessions) is measured by indicators Y1 through Y4. The Mplus programming language contains several defaults that are commonly implemented aspects of model specification (but nonetheless can be easily overridden by additional programming). For instance, Mplus automatically sets the first indicator after the "BY" keyword as the marker indicator (e.g., Y1) and freely estimates the factor loadings for the remaining indicators in the list (Y2–Y4). By default, all error variances are freely estimated and all error covariances and indicator cross-loadings are fixed to zero; the factor variances and covariances are also freely estimated by default. These and other convenient features in Mplus are very appealing to the experienced SEM researcher; however, novice users should become fully aware of these system defaults to ensure their models are specified as intended.

Selected results for the two-factor solution are presented in Table 16.1. Although not shown in Table 16.1, the Mplus results indicated that the two-factor model fit the data well, as evidenced by the goodness-of-fit statistics and other fit diagnostic information (modification indices, standardized residuals); for example, model χ^2 (19) = 21.73, p = .30. The unstandardized and completely standardized estimates can be found under the headings "MODEL RESULTS" and "STANDARDIZED MODEL RESULTS," respectively (partially standardized estimates have been deleted from the output). Starting with the completely standardized solution, the factor loadings can be interpreted along the lines of standardized regression coefficients in multiple regression. For instance, the factor loading estimate for Y1 is .645, which would be interpreted as indicating that a standardized unit increase in the Obsessions factor is associated with an .645 standardized score increase in Y1. However,

Table 16.1 Mplus Syntax and Selected Output for CFA Model of Obsessions and Worry

```
TITLE: TWO-FACTOR CFA OF OBSESSIONS AND WORRY
DATA:
  FILE IS CFA.DAT;
  TYPE IS STD CORR;
  NOBSERVATIONS = 400;
VARIABLE:
  NAMES   ARE   Y1-Y8;
ANALYSIS: ESTIMATOR=ML;
MODEL: OBS  BY  Y1-Y4;
       WORRY  BY  Y5-Y8;
OUTPUT: STANDARDIZED MODINDICES(4) RES;
```

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
OBS BY				
Y1	1.000	0.000	999.000	999.000
Y2	1.054	0.086	12.326	0.000
Y3	0.863	0.082	10.562	0.000
Y4	1.067	0.087	12.229	0.000
WORRY BY				
Y5	1.000	0.000	999.000	999.000
Y6	1.185	0.154	7.684	0.000
Y7	0.884	0.119	7.427	0.000
Y8	0.423	0.078	5.413	0.000
WORRY WITH				
OBS	0.629	0.167	3.757	0.000
Variances				
OBS	3.081	0.463	6.647	0.000
WORRY	1.694	0.321	5.281	0.000
Residual Variances				
Y1	4.326	0.357	12.110	0.000
Y2	1.754	0.211	8.307	0.000
Y3	3.269	0.269	12.139	0.000
Y4	2.028	0.228	8.908	0.000
Y5	2.764	0.283	9.777	0.000
Y6	2.567	0.332	7.736	0.000
Y7	2.974	0.269	11.069	0.000
Y8	2.071	0.157	13.180	0.000

(*continued*)

Table 16.1 (Continued)

STANDARDIZED MODEL RESULTS					
STDYX Standardization					
		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
OBS	BY				
Y1		0.645	0.035	18.419	0.000
Y2		0.813	0.026	30.759	0.000
Y3		0.642	0.035	18.266	0.000
Y4		0.796	0.027	29.310	0.000
WORRY	BY				
Y5		0.616	0.048	12.738	0.000
Y6		0.694	0.048	14.423	0.000
Y7		0.555	0.049	11.277	0.000
Y8		0.358	0.055	6.500	0.000
WORRY	WITH				
OBS		0.275	0.063	4.386	0.000
Variances					
OBS		1.000	0.000	999.000	999.000
WORRY		1.000	0.000	999.000	999.000
Residual Variances					
Y1		0.584	0.045	12.933	0.000
Y2		0.339	0.043	7.880	0.000
Y3		0.587	0.045	13.007	0.000
Y4		0.366	0.043	8.470	0.000
Y5		0.620	0.060	10.390	0.000
Y6		0.519	0.067	7.780	0.000
Y7		0.692	0.055	12.656	0.000
Y8		0.872	0.039	22.166	0.000
R-SQUARE					
Y1		0.416	0.045	9.210	0.000
Y2		0.661	0.043	15.379	0.000
Y3		0.413	0.045	9.133	0.000
Y4		0.634	0.043	14.655	0.000
Y5		0.380	0.060	6.369	0.000
Y6		0.481	0.067	7.211	0.000
Y7		0.308	0.055	5.639	0.000
Y8		0.128	0.039	3.250	0.001

because Y1 loads on only one factor, this estimate can also be interpreted as the correlation between Y1 and the Obsessions latent variable. Accordingly, squaring the factor loading provides the proportion of variance in the indicator that is explained by the factor; for example, 41.6 percent of the variance in Y1 is accounted for by Obsessions ($.645^2 = .416$). In the factor analysis literature, these estimates are referred to as *communalities*, the model-implied estimate of the amount of true score variance in the indicator. Communalities are provided in the Mplus output in Table 16.1 under the R-SQUARE heading.

The completely standardized estimates under the "Residual Variances" heading (see Table 16.1) represent the proportion of variance in the indicators that has not been explained by the latent variables (i.e., unique variance). For example, these results indicate that 58.4 percent of the variance in Y1 was not accounted for by the Obsessions factor. Note that the analyst could readily hand-calculate these estimates by subtracting the indicator communality from one; for example, $1 - .645^2 = .584$. The completely standardized results also provide the correlation of the Obsessions and Worry factors (WORRY WITH OBS = .275). Although statistically significant, the magnitude of this correlation would suggest that the constructs of Obsessions and Worry possess acceptable discriminant validity. If this correlation had been higher (e.g., $r = .85$ or above; Kenny, 1979), this would raise concerns about the uniqueness of the two constructs and might suggest that the two factors should be collapsed into a single dimension.

The first portion of the Mplus results shown in Table 16.1 is the unstandardized solution (under the heading "MODEL RESULTS"). In addition to each unstandardized estimate (under "Estimate" heading), Mplus provides the standard error of the estimate ("S.E."); the test ratio, which can be interpreted as a z statistic ("Est./S.E."; i.e., values greater than 1.96 are significant at $\alpha = .05$, two-tailed), and the exact two-sided probability value. In more recent versions of this software program, Mplus also provides standard errors and test statistics for completely standardized estimates (also shown in Table 16.1, and discussed above), as well as partially standardized estimates (which have not been included in Table 16.1). The standard errors and significance tests for the unstandardized factor loadings for Y1 and Y5 are unavailable because these variables were used as marker indicators for the Obsessions and Compulsions factors, respectively (i.e., their unstandardized loadings were fixed to 1.0). The variances for the Obsessions and Worry latent variables are 3.08 and 1.69, respectively. These estimates can be calculated using the sample variances of the marker indicators multiplied by their respective communalities. As noted earlier, the communality for Y1 was .416. Thus, 41.6 percent of the sample variance of Y1 ($SD^2 = 2.725^2 = 7.426$; cf. Table 16.1) is passed along as variance of the Obsessions latent variable; 7.426(.416) = 3.08. As in the completely standardized solution, the factor loadings are regression coefficients expressing the direct effects of the latent variables on the indicators, but in the unstandardized metric (e.g., a unit increase in Obsessions is associated with a 1.054 increase in Y2). The WORRY WITH OBS estimate (0.629) is the factor covariance of Obsessions and Worry. The residual variances are the indicator uniquenesses or errors (i.e., variance in the indicators that was not explained by the Obsessions and Worry latent variables).

Use of CFA for Disorder Nosology

Measurement models along the lines of the model described above can be used to inform the classification of mental disorders. For example, numerous studies have used CFA to examine the latent structure of posttraumatic stress disorder (PTSD), given that PTSD is thought to be multidimensional and defined in the DSM by multiple criteria sets (i.e., reexperiencing, avoidance and numbing, and hyperarousal symptoms). Palmieri, Weathers, Difede, and King (2007) evaluated competing models of the structure of PTSD and found (along with other researchers) that two four-factor models of PTSD provided the best fit to the data, relative to one-, two-, and three-factor models. The model testing sequence included an explicit evaluation of the DSM-IV structure of PTSD, which presumes that the 17 symptoms reflect the constructs of reexperiencing, avoidance and numbing, and hyperarousal. CFA demonstrated that this DSM-IV-oriented three-factor model did not fit the data as well as two four-factor models; one four-factor model included the addition of a distinct dysphoria factor and was the prevailing model in the self-report data that were evaluated, and the second four-factor model included separate avoidance and emotional numbing factors and was the best-fitting model in the clinician rating data. The findings raise questions about the organization of the PTSD symptoms in the DSM-IV. The results of this study, and others like it (e.g., Baschnagel, O'Connor, Colder, & Hawk, 2005; Elklit & Shevlin, 2007; King, Leskin, King, & Weathers, 1998; Simms, Watson,

& Doebbeling, 2002), have led to the current proposal to revise the diagnosis in DSM-5 whereby the three-symptom-cluster definition is replaced by four symptom clusters that separate the avoidance criteria from features reflecting negative affect, mood, and cognition.

In addition to informing the nosology of a single diagnosis (e.g., PTSD), CFA can be used in the broader evaluation of the structure of psychopathology across diagnostic constructs. For instance, CFA has been used to evaluate models of common mental disorders that posit that the comorbidity of the unipolar mood, anxiety, and somatization disorders is due to an underlying common factor termed Internalizing, whereas the comorbidity of the substance use disorders and antisocial personality disorder is accounted for by a latent variable termed Externalizing (Kessler, Chiu, Demler, Merikangas, & Walters, 2005; Krueger, 1999; Krueger, Caspi, Moffitt, & Silva, 1998; Miller, Fogler, Wolf, Kaloupek, & Keane, 2008; Slade & Watson, 2006; Vollebergh et al., 2001). This latent structure is highly replicable across studies and samples (Krueger, Chentsova-Dutton, Markon, Goldberg, & Ormel, 2003), is stable over time (Krueger et al., 1998), and has a genetic basis (Wolf et al., 2010).

As a result of this factor analytic work, some have argued that the internalizing and externalizing constructs should be adopted by the DSM-5 as a means to reorganize the nosology and to demonstrate the relatedness and dimensionality of common disorders (cf. Goldberg, Krueger, & Andrews, 2009; Krueger & South, 2009). These latent variables have also been conceptualized as endophenotypes for use in the evaluation of genes and other biomarkers of psychopathology because these constructs are thought to more accurately reflect the end product of the gene than individual DSM-IV-defined disorders. Initial work using latent phenotypes in the search for genes associated with psychopathology has demonstrated the utility of this approach (Hettema et al., 2008).

CFA can be useful in the classification of disorders by evaluating the extent to which a proposed factor structure is the same across groups or subgroups of individuals. For example, before adopting a revised structure of mental disorders for DSM, investigators should demonstrate that the structure is consistent across gender, age, ethnicity, race, and other salient population subgroups. Although a thorough discussion is beyond the scope of the chapter, CFA lends itself well to answering questions of this sort. In particular, researchers can evaluate if a measurement model is invariant across groups of individuals using a multiple-group CFA design (cf. Brown, 2006). This involves imposing successively restrictive equality constraints on the parameters of the measurement models in two or more groups and determining if these constraints result in significant degradations in model fit (for applied examples, see Campbell-Sills, Liverant, & Brown, 2004; Kramer, Krueger, & Hicks, 2008; Krueger et al., 2003).

Higher-Order CFA

CFA can be used to study the higher-order structure of psychological disorders. For example, Markon (2010) used higher-order CFA to evaluate both the lower-order and higher-order structure of Axis I and II disorders. He demonstrated that symptoms of Axis I and II disorders loaded on 20 different lower-order factors that were fairly narrow in scope. In turn, these 20 factors loaded on one or more of four higher-order factors. These higher-order factors were broader than the lower-order factors and included the internalizing and externalizing dimensions.

A hypothesized example of this type of structure is shown in Figure 16.2. This type of structure implies that the correlations among the lower-order (or first-order) factors in the model are accounted for by the higher-order factor on which they load; for instance, in Figure 16.2, the variance and covariance of Depression, Generalized Anxiety Disorder (GAD), and Somatoform are accounted for by the higher-order Internalizing factor. The higher-order factor accounts for these associations in a manner that is more parsimonious than an intercorrelated first-order model. For example, in Figure 16.2, 15 factor correlations would be freely estimated if only the lower-order portion of the model was specified. However, the higher-order portion of the model reproduces these 15 associations with only seven freely estimated parameters (six loadings on the higher-order factors and one factor correlation between the two higher-order factors). The lower-order factors can be thought of as indicators of the higher-order factor, and standard model identification rules, as described previously in this chapter, must be met in order for the higher-order portion of the model to be identified. This type of model also implies that the association between the latent higher-order variable (e.g., Internalizing in Fig. 16.2) and an observed indicator (e.g., the first measure of Depression in Fig. 16.2) is a mediated one, with the lower-order factor (Depression) serving as the intervening variable. The direct influence of

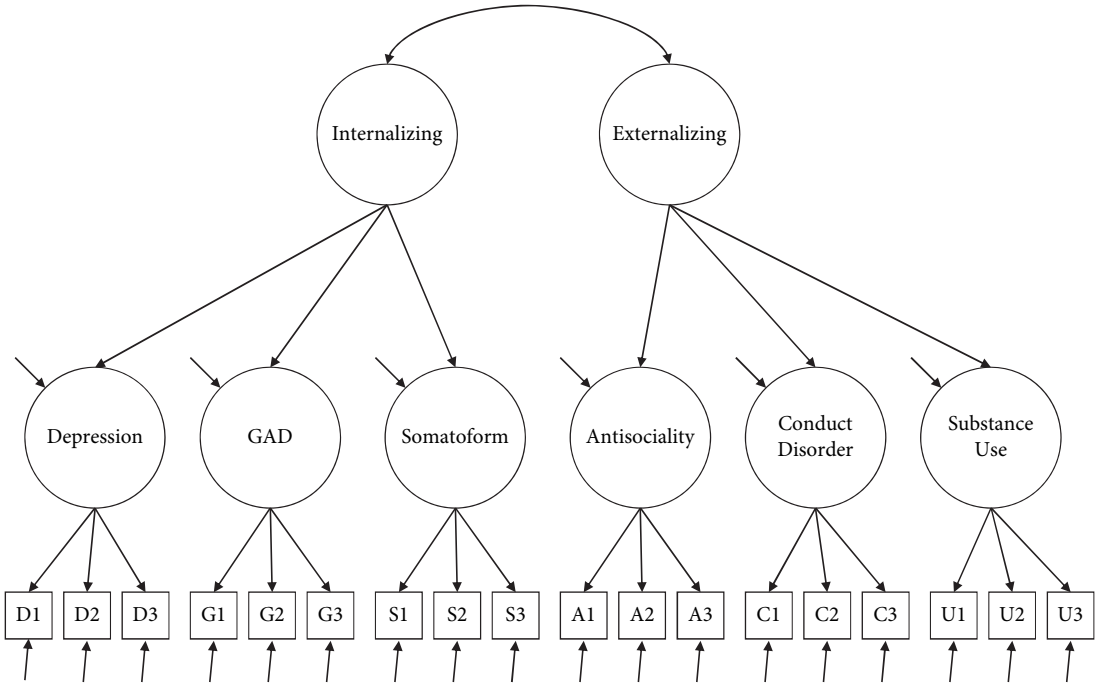


Figure 16.2 Higher-order confirmatory factor analysis.

the higher-order variable on an individual indicator in the model can be computed by taking the product of the path from the higher-order factor to the lower-order factor and the path from the lower-order factor to the observed indicator (i.e., in Fig. 16.2, the product of the path from Internalizing to Depression and the path from Depression to the first measure of Depression).

Higher-order models hold many conceptual advantages, such as in the classification of mental disorders. Specifically, these models provide a way to think about the various levels for representing psychopathology, as each level may be appropriate for different purposes. For example, returning to the example in Figure 16.2, if a researcher is interested in identifying biomarkers for psychopathology, they may want to consider using the highest level of the model (i.e., the Internalizing variable) because this broad factor may be more strongly related to biological mechanisms in contrast to more narrow constructs (i.e., it is unlikely that there are distinct biological pathways for DSM-defined depression vs. generalized anxiety). The cost and time burden associated with having multiple measures of each indicator of Internalizing might be worthwhile to the researcher if the use of the Internalizing measurement model allows the researcher to observe associations with biomarkers that would not otherwise be observed. In contrast, a researcher evaluating the effectiveness of a psychological treatment may want to use the lower-level factors in the model (i.e., in Fig. 16.2, the latent variable of Depression), as this level of the model provides greater specificity for the construct of depression while still improving upon its measurement relative to a single-indicator approach. Finally, a researcher who is not focused on depression but simply wants to include an inexpensive and quick measure of the construct might choose to simply use a single indicator of depression that has been shown to have a strong association with the latent construct.

Multitrait–Multimethod Matrices

Another application of CFA for the study of psychopathology is construct validation. Although a simple CFA is, itself, a form of construct validation, it is not the most comprehensive test of a construct. Rather, a stronger test is the analysis of a multitrait–multimethod matrix (MTMM; Campbell & Fiske, 1959). For example, suppose a researcher wants to demonstrate the validity of the construct of depression. A stringent test of this would involve collecting indicators of depression from multiple assessment modalities (e.g., a self-report measure, an interview, and a collateral rating measure) and collecting multiple indicators of other constructs that are thought

to be strongly versus weakly related to depression. For example, measures of anxiety and psychopathy (each assessed via self-report, interview, and collateral report) could be used to assess the convergent and discriminant validity of the depression construct. In one CFA approach to MTMM (*correlated method factors*), each indicator of a psychopathology construct is specified to load on a latent variable of psychopathology (a trait factor), and each indicator that shares an assessment modality would also be specified to load on a method factor (e.g., all interview measures would load on an interview method factor). Figure 16.3 depicts a correlated method factors CFA. This model, although shown in Figure 16.3 because of its conceptual appeal, tends to have problems with model convergence (Brown, 2006; Kenny & Kashy, 1992). Thus, a more commonly used CFA approach to MTMM data does not entail the specification of method factors; instead, the residual variances of indicators from the same assessment modality are specified to freely intercorrelate with one another. This approach is termed the *correlated uniqueness* model, and it sidesteps the potential model identification and nonconvergence problems that often arise with the correlated methods factors approach to MTMM. In the correlated uniqueness model, evidence for convergent validity is found when the indicators thought to measure the same construct load strongly and significantly on the appropriate trait factor (i.e., after accounting for method effects). However, if the trait factors correlate with one another substantially, this would suggest poor discriminant validity. Small to moderate correlated errors among indicators from the same assessment modality would indicate the presence of method effects and would not be considered problematic unless the strength of their association was large in the context of weak factor loadings on the trait factors. The latter pattern of results would suggest that the scores on the measures are unduly influenced by the modality of the assessment as opposed to the substantive construct of interest.

An applied example of the CFA approach to MTMM data can be found in a study by Kollman, Brown, and Barlow (2009) in which the construct of acceptance was evaluated. Specifically, the authors used three forms of self-report items of the construct of acceptance and compared them to three types of self-report items tapping cognitive reappraisal and perceived emotional control. The CFA revealed that the acceptance items loaded on the acceptance latent variable and were distinct from the cognitive reappraisal, perceived emotional control, and method factors, but that the acceptance latent variable did not evidence the hypothesized associations with other measures assessing psychopathology and well-being.

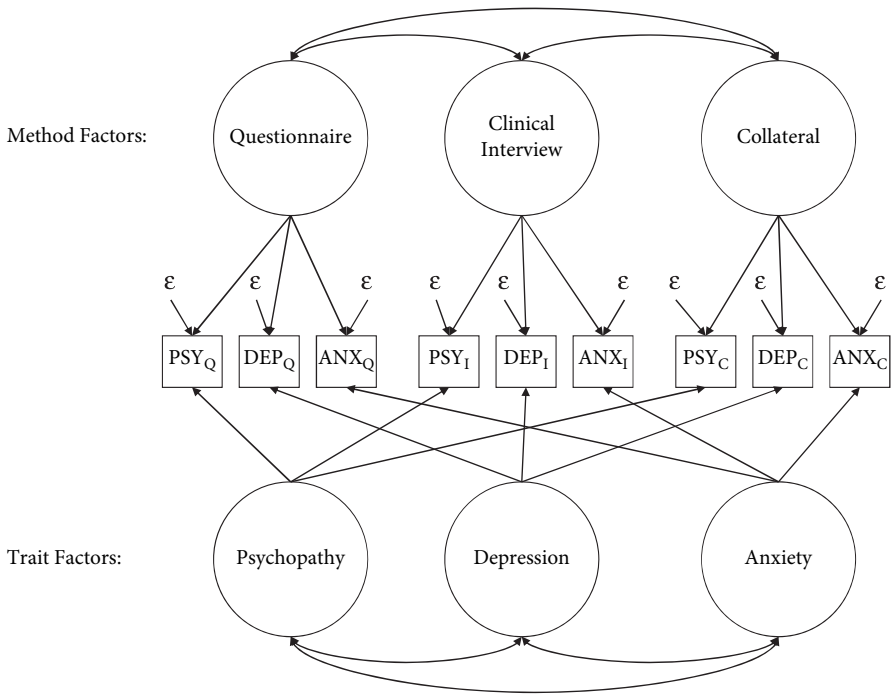


Figure 16.3 Multitrait–multimethod confirmatory factor analysis.

These results were viewed as providing support for the uniqueness of the acceptance construct but suggested that, despite its construct validity, the construct may be less useful clinically. This type of MTMM is an elegant method for evaluating construct validity and the clinical utility of constructs and highlights one of the many strengths of CFA.

Use of CFA to Study the Etiology of Psychopathology

CFA can be used to investigate the genetic and environmental risk of one or more psychological disorders. In particular, common forms of twin modeling are based on a CFA in which the observed variance of a disorder is estimated to be a function of genes, the shared environment (i.e., aspects of the environment that are common across members of a twin pair), the nonshared environment (i.e., aspects of the environment not shared across twin pairs), and measurement error (in the case of twin models using observed indicators, measurement error is confounded with variance attributable to the nonshared environment). The genetic and environmental factors are unobserved and are modeled as latent variables. The strength of the genetic and environmental paths on the observed indicators can be solved because of known associations among the latent genetic and environmental factors for members of a twin pair. These associations are specified in the CFA as constraints on the model, and these constraints differ for monozygotic (i.e., identical) versus dizygotic (i.e., fraternal) twin pairs. To evaluate these paths, separate genetic and environmental factors are modeled for each member of a twin pair. The correlation between genetic factors is constrained to 1 for monozygotic twins because these twins share 100 percent of their genes; in contrast, this association is constrained to .50 for dizygotic twins because such twins share, on average, 50 percent of their genes. The correlation between the shared environmental factors for each twin pair is set to 1 for both monozygotic and dizygotic twins, while the correlation between the nonshared environment factors is constrained to 0 across members of a twin pair. In other words, solving for the genetic and environmental paths in the model involves use of a multiple-group design (with twin type as the grouping variable), with some group-specific constraints on the associations between the latent genetic and environmental variables. This type of CFA has many variations that allow investigators to evaluate the necessity of genetic and environmental paths and to evaluate the extent to which the same genetic and environmental factors might contribute to more than one observed indicator. This approach contributes meaningfully to studying the etiology of psychopathology.

Structural Regression Models: Directional Relationships Among Latent Variables

We now consider the use of structural regression models of psychopathology to inform our understanding of the correlates, predictors, and effects of latent variables of psychopathology. These types of SEM models are referred to as "structural regression models" because directional relationships are specified among the variables (e.g., unlike CFA, where the latent variables are only specified to be freely intercorrelated, see Fig. 16.1). In addition to estimating the structural relationships among latent variables, the SEM model can include variables that are represented by a single indicator. SEM can also evaluate interaction effects of latent variables (i.e., latent moderation). Thus, SEM provides a flexible approach for evaluating many different types of variables and relationships among variables. In this section, we provide examples of the types of questions that can be evaluated using structural regression models.

Before specifying a structural regression model, the researcher should evaluate the acceptability of a measurement model (CFA). Poor-fitting SEM models usually stem from problems with the measurement model (i.e., there are more potential sources of ill fit in a measurement model than in a structural model). Thus, an acceptable measurement model should be developed (i.e., in which the latent variables are allowed to freely intercorrelate) prior to moving on to a structural regression model, which often places constraints on the nature of the relationships among the latent variables (e.g., in a full mediational model, a latent variable may exert an influence on a downstream latent variable only through an intervening latent variable).

Examples of Structural Regression Models

Consider a straightforward structural regression model involving three latent exogenous (independent) variables and two latent endogenous (dependent) variables. As seen in Figure 16.4, the three exogenous factors are Depression, Anxiety, and Psychopathy, and the two endogenous factors are Social Alienation and General Distress. For presentational clarity, the measurement portion of the model is not shown in Figure 16.4 (e.g., indicators, factor loadings). Whereas the three exogenous variables are

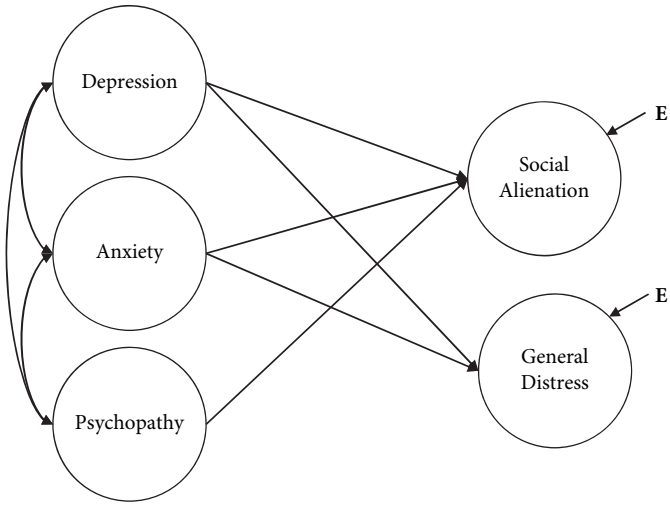


Figure 16.4 Structural regression model.

expected to contribute significantly to the prediction of Social Alienation, only Anxiety and Depression are expected to be predictive of General Distress. The exogenous variables are specified to be freely intercorrelated (double-headed curved arrows), and thus the paths emanating from these factors would be interpreted as partial regression coefficients (e.g., the Depression → Social Alienation path is holding Anxiety and Psychopathy constant). The model is not structurally just-identified because all five latent variables are not directly associated with one another by either regression paths (e.g., Depression → Social Alienation) or correlations (e.g., Depression with Anxiety). Specifically, the model specified in Figure 16.4 assumes that Psychopathy does not have a significant direct effect on General Distress. In addition, the residual variances of the two endogenous variables (depicted as E's in Fig. 16.4) are presumed to be uncorrelated (i.e., there is no double-headed, curved arrow connecting the residual variances). This model specification anticipates that any covariance that exists between Social Alienation and General Distress will be accounted for by the exogenous factors. As noted earlier, because poor model fit may arise from misspecification of the structural portion of the model (e.g., Psychopathy may have a direct effect on General Distress), testing this model should be preceded by establishing an acceptable five-factor measurement model (CFA).

We can see an applied example of the structural regression model in a study by Brown, Chorpita, and Barlow (1998), who found that virtually all the considerable covariance among latent variables corresponding to the DSM-IV constructs of unipolar depression, social phobia, generalized anxiety disorder, obsessive-compulsive disorder, and panic disorder/agoraphobia was explained by the higher-order dimensions of negative affect and positive affect. Although the results were consistent with the notion of neuroticism/negative affect as a broadly relevant dimension of vulnerability, the results indicated that the DSM-IV disorders were differentially related to negative affect, with the Depression and Generalized Anxiety Disorder factors evidencing the strongest associations. In accord with a reformulated hierarchical model of anxiety and depression (Mineka, Watson, & Clark, 1998), positive affect was predictive of Depression and Social Phobia only. Moreover, the results indicated that although a latent variable of Autonomic Arousal was strongly related to a Panic Disorder/Agoraphobia latent variable, it was not relevant to other DSM-IV anxiety disorder constructs such as Social Phobia and Obsessive-Compulsive Disorder.

Another common type of model found in the psychopathology literature is the *mediation model*. SEM allows for an elegant method to evaluate such models. For example, suppose a researcher wanted to evaluate if inflammation serves to mediate the association between depression and cardiac disease. Perhaps the researcher has multiple indicators of inflammation (i.e., values of various peripheral cytokine markers), multiple indicators of depression symptoms (i.e., several questionnaires), and one continuous clinical severity rating of cardiac disease (Fig. 16.5). Note that unlike the prior SEM, this model is structurally saturated, because all latent variables are connected to one another by direct effects. Thus, the structural portion of the model is just-identified. However, the model has positive *df*

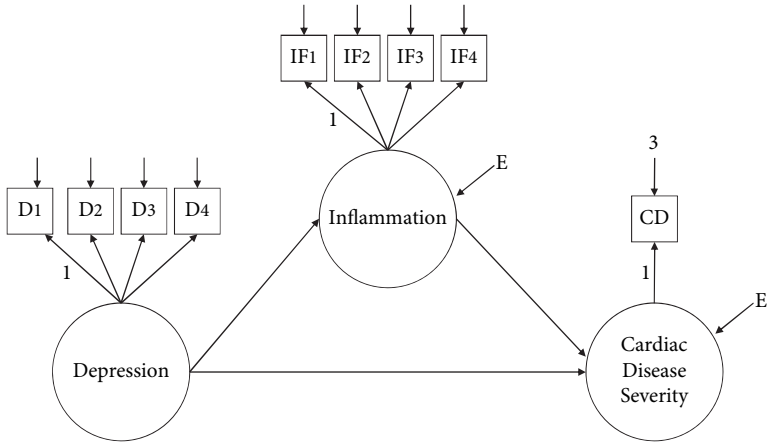


Figure 16.5 Mediational model with a single-indicator outcome.

because of the measurement portion of the model, which is overidentified (i.e., the goodness of fit of the measurement model and structural model are identical). For the SEM model, the researcher intends to create latent variables of depression and inflammation and to use the cardiac disease severity variable as a single indicator that has been adjusted for measurement error. The estimate of measurement error is based on a published report of the interrater reliability of the cardiac disease measure. To set the measurement error variance of the cardiac disease variable, one takes the product of the sample variance for the measure with an estimate of the unreliability of the measure: VAR(X)(1 – ρ), where X is the clinical severity rating of cardiac disease and ρ is the reliability estimate of this measure. In this example, if the reliability estimate (e.g., an intraclass correlation coefficient) for the measure is .80 and the sample variance of the measure is 15.0, the error variance would be 3 [i.e., (15)(1 – .8)]. The error variance for the cardiac disease severity variable would then be fixed to a value of 3.0 (cf. Fig. 16.5) in the program syntax.

If requested in the program syntax, the latent variable software program such as Mplus will compute the size and statistical significance of the indirect effect (i.e., Depression → Inflammation → Cardiac Disease Severity). Although the size and significance of this particular indirect effect can be readily hand-calculated using the delta method (cf. MacKinnon, 2008), the advantages of this program feature are more pronounced in other contexts (e.g., when the indirect effects of interest emanate from a more complex causal chain).

An applied example of a latent mediation model can be found in a study by Miller, Vogt, Mozley, Kaloupek, and Keane (2006), who evaluated whether latent temperament variables mediated the association between PTSD and comorbid substance use (both disorder constructs were modeled as separate latent variables). Specifically, the temperament variable "disconstraint" (characterized by impulsivity, recklessness, and nontraditionalism) mediated the association between PTSD and alcohol and drug use problems while the temperament variable "negative emotionality" mediated the association between PTSD and alcohol problems. The findings suggest important differences in the correlates of alcohol versus drug use problems that co-occur in PTSD and have implications for the conceptualization of and development of treatments for these comorbidities. SEM provided an elegant and parsimonious method for evaluating these associations.

Newer Applications of SEM

One relatively new use of SEM in psychopathology research is the evaluation of relationships among brain regions. Specifically, researchers using magnetic resonance imaging (MRI) have used exploratory factor analysis to identify brain regions that are physically related to one another as a function of brain volume (i.e., brain volumes for different brain regions served as the observed indicators). The hypothesized relationships between latent brain structures or systems are then evaluated using structural path modeling (i.e., Yeh et al., 2010). This approach has been used to evaluate differences in structural relationships among brain regions across patient groups (Yeh et al., 2010) and clearly reflects a sophisticated approach to dealing with the large amounts of data generated by MRI research. The latent variable approach holds the advantage of reducing the risk

of type I error by requiring fewer analyses than if all brain regions were evaluated independently.

Another novel use of SEM is research evaluating the association between genes and psychological symptoms or other health-behavior problems. For example, one study used single nucleotide polymorphisms (SNPs) from the same gene as observed indicators of genetic information (Nock et al., 2009). That is, genes were represented as latent variables in SEM for the evaluation of the association between the latent genes and metabolic disease. Further, this study demonstrated the benefits of using the SEM approach over univariate analytic approaches by also evaluating the same types of associations using univariate models. This work represents the merger of state-of-the-art genotyping technology with state-of-the-art multivariate data analytic techniques. Although individual SNPs on a single gene typically explain very small portions of the variance in complex diseases and traits, this latent variable approach is likely to yield stronger and more robust genetic effects.

A Caveat About Those Arrows

One important point to remember is that SEM has no more ability to determine causation than does simple correlation. Despite the directional arrows and the notion of "causal" path modeling, researchers' ability to make causal determinations about the associations in their data is a function of research design, not analytic design. One cannot infer causal associations in the context of a cross-sectional study without, for example, treatment randomization simply by using SEM as the analytic approach. That is to say, although the researcher may have a good reason to believe that variable #1 is an X variable and variable #2 is a Y variable, ultimately he or she must take into consideration that the directional nature of associations between these two variables may be reversed (or that omitted variables may be responsible for the observed relationships among the study variables). More broadly, this point relates to the fact that in evaluating SEMs, there are usually alternative models of the associations among the variables of interest that would fit the data as well as (or better) than the hypothesized model. Researchers must always be mindful of these issues when interpreting the results of SEM studies.

Latent Growth Curve Modeling

SEM also provides a very powerful and flexible framework for analyzing time-series data. Although there are many approaches to the statistical analysis of repeated measures, the two most popular SEM-based methods are the *autoregressive model* (AR) and the *latent growth curve model* (LGM). This chapter will focus primarily on LGM, but a brief discussion of AR models is presented here to foster the comparison of these two SEM-based approaches.

Prior to the advent of LGM (e.g., Meredith & Tisak, 1990), the AR model was the most common SEM method for analyzing longitudinal data. Path diagrams for two AR models are presented in Figure 16.6. Although the path diagrams depict single indicators, AR models can be readily extended to latent variable outcomes (cf. Farrell, 1994), along the lines of the structural regression models discussed earlier in this chapter. The first path diagram in Figure 16.6 is a univariate AR model in which a single measure (a depression scale) is collected at three time points. As shown in Figure 16.6, a measure at a subsequent time point is regressed onto the same measure at the previous time point (i.e., the *autoregressive* path, β). Thus, change in the construct is an additive function of the immediately preceding measure (β) and a random disturbance (ε, variance in the subsequent measure not explained by the preceding measure). Moreover, the AR model assumes that a variable collected at an earlier time point has a unique effect only on the variable that immediately follows it. For instance, in the univariate AR model in Figure 16.6, the influence of

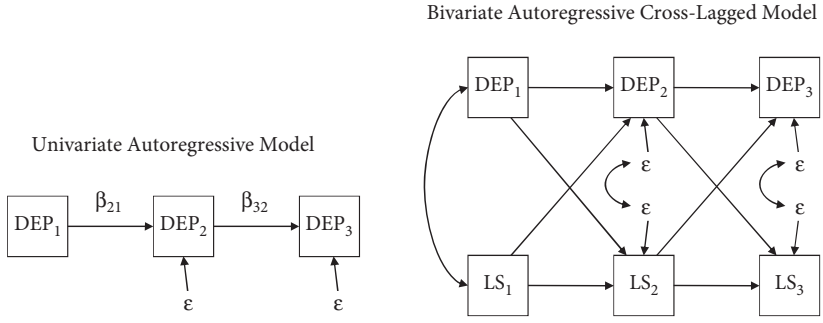


Figure 16.6 Univariate and bivariate autoregressive models. DEP = depression, LS = life stress.

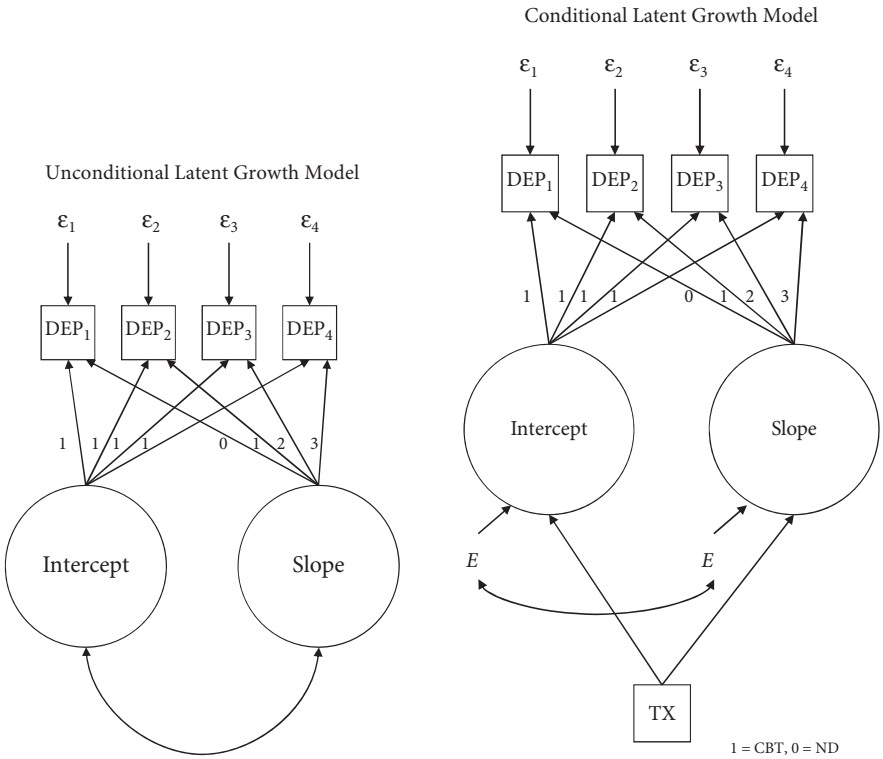
DEP_1 on DEP_3 is fully mediated by DEP_2 (i.e., the correlation of DEP_1 and DEP_3 is reduced to zero when controlling for DEP_2). The AR model is often extended to the analysis of two or more variables. The second path diagram in Figure 16.6 is a bivariate AR model in which both depression and life stress are measured on three testing occasions. The bivariate AR model specification includes both autoregressive paths and cross-lagged paths. Accordingly, multivariate AR models are often referred to as *autoregressive cross-lagged models*. The cross-lagged paths are estimates of whether a variable from the immediately preceding time point explains unique variance in the other variable (e.g., $LS_1 \rightarrow DEP_2$), holding the autoregressive path constant (e.g., $DEP_1 \rightarrow DEP_2$). In multiple-wave panel data, subsequent time points (e.g., from Time 2 to Time 3) are used to further clarify the nature of the temporal cross-lagged relationships (e.g., replication across waves, evidence of a bidirectional influence). As shown in Figure 16.6, the disturbances (residual variances) for the two variables from a given assessment are specified to be correlated in part because the paths from the preceding time point may not explain all the covariance of these variables (e.g., the covariance of DEP_2 and LS_2 may not be fully accounted for by the paths emanating from DEP_1 and LS_1).

Although there are scenarios where AR models are appropriate (discussed later), a potentially limiting factor is that this approach only renders fixed effects among variables at time-adjacent intervals (e.g., assumes that the autoregressive and cross-lagged effects are the same for all individuals in the sample). Unlike AR models, LGMs estimate the rate, shape, and variability of change in outcome measures over time. Thus, LGMs are more appropriate than AR models in situations where the researcher is interested in evaluating the amount of and individual differences in change (e.g., treatment outcome studies, developmental studies). A growth trajectory is fit to the repeated measures data of each participant in the sample. Thereafter, the LGM produces *fixed-effect* as well as *random-effect* estimates from these trajectories. For instance, one LGM fixed effect expresses how much participants in the sample changed on average; the corresponding random effect represents variability around this average (i.e., individual differences in change).

Unconditional LGM

Figure 16.7 presents path diagrams for two LGMs as well as data that will be used as input for the examples to follow. The example uses a simulated sample of 200 participants who underwent a 12-week, randomized controlled trial for depression (100 cases received cognitive-behavioral treatment [CBT], 100 were assigned to nondirective treatment). A questionnaire measure of depression was administered on four occasions during the active treatment phase (every 3 weeks).

The first path diagram in Figure 16.7 is an *unconditional LGM*. It is so called because this LGM estimates the growth trajectories but does not include covariates that may affect the growth trajectories (e.g., variables that may explain individual differences in change). As reflected by this path diagram, in essence LGM is an adaptation of the confirmatory factor analytic model where the unstandardized factor loadings are prespecified by the researcher in accord with predictions about the shape of temporal change. Thus, Time in the LGM is represented by a factor loading matrix. In the basic unconditional LGM, two growth factors are specified: an Intercept and a Slope. If the model is specified correctly, the growth factors will sufficiently reproduce the sample variance–covariance matrix as well as the sample means. Unlike many other applications of SEM, the indicator means are always included in LGM because the analysis is estimating the rate and shape of temporal change. As in multiple regression, the Intercept is a constant. Accordingly, the unstandardized factor loadings emanating from the Intercept to the repeated observed measures are always fixed to 1.0 (the Intercept factor equally influences each repeated measure). How the Slope factor loadings are specified depends on the anticipated shape of the growth trajectory as well as the timing of the repeated assessments. In this example, the assessment waves were spaced equally (every 3 weeks) and linear growth was expected. If linear growth is tenable, then the differences between the observed means of the outcome variable at time-adjacent waves should be roughly the same, except for sampling error (e.g., considering the sample means in Fig. 16.7, the change from DEP_1 to DEP_2 is roughly the same degree as the change in DEP_2 to DEP_3). Thus, in this example, the unstandardized Slope factor loadings were fixed to values of 0, 1, 2, and 3 for the four respective repeated measures. The LGM would be specified differently if nonlinear change was expected (discussed later), or if linear change was expected but the repeated measures were collected at unequal intervals (e.g., Slope factor loadings of 0, 1, 4, and 5 would capture linear change in an outcome assessed at Weeks 1, 2, 5, and 6). In addition to representing the anticipated form of the growth trajectories, this Slope factor loading specification centers the Intercept on the first time




```
Input Data

Correlations:
             DEP1     DEP2     DEP3     DEP4     TX

DEP1         1.000
DEP2         0.679    1.000
DEP3         0.627    0.687    1.000
DEP4         0.577    0.659    0.723    1.000
TX          -0.056   -0.143   -0.245   -0.276    1.000

Means:       34.573   31.458   27.625   23.803   0.500
SDs:         18.752   18.658   19.163   19.821   0.577
```

Figure 16.7 Unconditional and conditional latent growth models. DEP = depression, TX = treatment condition (CBT = cognitive-behavioral treatment, ND = nondirective treatment).

point. One can view the Slope factor loadings as values of Time. By fixing the first Slope factor loading to zero, the Intercept provides the model-implied estimates of the outcome variable when Time = 0 (e.g., the predicted mean and variance of DEP_1). Although the LGM is usually centered on the first time point, it can be centered on any other observation if substantively useful. For instance, Slope factor loadings of –3, –2, –1, and 0 would center the Intercept on the last observation (i.e., model-implied estimates of the mean and variance of DEP_4). The respecification of the Slope factor loadings in this fashion does not alter the goodness of fit of the model (relative to centering on the first time point) because it is an equivalent parameterization of the unconditional LGM.

Model identification and estimation. Table 16.2 presents the Mplus 6.1 syntax for estimating the unconditional LGM. On the first two lines of the "MODEL:" command, the Intercept (INT) and Slope (SLP) growth factors are specified and the factor loadings of the four repeated measures are fixed in the fashion discussed in the preceding section. The third line of the "MODEL:" command fixes the intercepts of the four repeated measures to zero (e.g., [DEP1@0]). This is a mandatory specification in LGM as it allows the indicator means to be predicted by the growth factors. The

Table 16.2 Mplus Syntax and Selected Output for the Unconditional Latent Growth Model

```
TITLE: UNCONDITIONAL LATENT GROWTH MODEL OF DEPRESSION
DATA:
  FILE IS LGM.DAT;
  TYPE IS MEANS STD CORR;
  NOBS = 200;
VARIABLE:
  NAMES ARE DEP1 DEP2 DEP3 DEP4 TX;
  USEV = DEP1 DEP2 DEP3 DEP4;
ANALYSIS: ESTIMATOR = ML;
MODEL:  INT BY DEP1@1 DEP2@1 DEP3@1 DEP4@1;
          SLP BY DEP1@0 DEP2@1 DEP3@2 DEP4@3;
          [DEP1@0 DEP2@0 DEP3@0 DEP4@0 INT SLP];
OUTPUT:   SAMPSTAT STANDARDIZED MODINDICES(4) TECH1;
```

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
INT	BY				
DEP1		1.000	0.000	999.000	999.000
DEP2		1.000	0.000	999.000	999.000
DEP3		1.000	0.000	999.000	999.000
DEP4		1.000	0.000	999.000	999.000
SLP	BY				
DEP1		0.000	0.000	999.000	999.000
DEP2		1.000	0.000	999.000	999.000
DEP3		2.000	0.000	999.000	999.000
DEP4		3.000	0.000	999.000	999.000
SLP	WITH				
INT		-10.825	9.133	-1.185	0.236
Means					
INT		34.789	1.265	27.503	0.000
SLP		-3.620	0.406	-8.911	0.000
Intercepts					
DEP1		0.000	0.000	999.000	999.000
DEP2		0.000	0.000	999.000	999.000
DEP3		0.000	0.000	999.000	999.000
DEP4		0.000	0.000	999.000	999.000
Variances					
INT		246.995	33.573	7.357	0.000
SLP		13.305	4.428	3.005	0.003

(*continued*)

Table 16.2 (Continued)

Residual Variances				
DEP1	104.784	20.658	5.072	0.000
DEP2	106.041	14.104	7.518	0.000
DEP3	106.089	14.354	7.391	0.000
DEP4	91.079	20.459	4.452	0.000

third line also freely estimates the means on the Intercept and Slope factors, [INT SLP]. The variances and covariance of the Intercept and Slope are freely estimated by Mplus default, as are the residual variances of the repeated measures (although not illustrated in the current example, the residual variances of the indicators are often constrained to equality in the LGM, analogous to HLM estimation; cf. Raudenbush & Bryk, 2002). Thus, there are nine freely estimated parameters in this model: the means and variances of the Intercept and Slope (four estimates in total), the covariance of the Intercept and Slope, and the four indicator residual variances. There are 14 pieces of information used as input: four indicator variances, six indicator covariances, and four indicator means (see unconditional LGM in Fig. 16.7). Thus, this LGM is overidentified with $df = 5$ (14 – 9). It should be noted here that LGM requires a bare minimum of three repeated observations (e.g., in addition to issues of statistical underidentification, the use of only two time points would not allow for evaluation of the shape of the growth trajectory). Although an LGM with three time points can be estimated, it is bordering on being underidentified. For instance, an unconditional LGM that estimates linear growth underlying three repeated measures is overidentified, with just a single degree of freedom (nine elements of the input data, eight freely estimated parameters). If a poor-fitting model is encountered, this leaves little room for alternative model specifications (e.g., nonlinear growth). The availability of several repeated observations not only assists in the statistical identification and modeling flexibility of the LGM, but fosters the statistical power of the analysis (cf. Muthén & Curran, 1997).

Interpretation. The unconditional LGM model shown in Figure 16.7 fit the data well; for example, $\chi^2(5) = 0.41$, $p = .99$. If poor fit was encountered, there are two areas of the model that could be revised, if justified: the Slope factor loadings and the covariances of the indicator residuals. Other aspects of the LGM specification should not be altered (e.g., Intercept factor loadings should always be fixed to 1.0; the intercepts of the repeated observed measures should always be fixed to 0). As noted earlier, the Slope factor loadings were fixed to specific values (i.e., 0, 1, 2, and 3) based on the prediction that patients' symptom reduction during the active treatment phase would be linear. If the growth trajectory was nonlinear, then the model would not fit the data satisfactorily and the Slope factor loadings should be specified in a different fashion (discussed below). In the current model, the indicator residual variances were specified to be uncorrelated (i.e., by Mplus default, these error covariances are fixed to 0). This specification is based on the assumption that the relationships among the repeated measures of depression observed in the sample data are due solely to the underlying growth process (i.e., the Intercept and Slope growth factors account for all of the temporal covariance of the indicators). Although traditional analytic approaches (e.g., analysis of variance) assume that the error variances of repeated measures are equal and independent, these assumptions are often violated in real datasets. Thus, one of the many advantages of LGM compared to other approaches is the ability to formally test these assumptions and to incorporate an error theory in the model (e.g., temporal equality of error variances, independence of errors).

Table 16.2 presents the unstandardized estimates from the Mplus output for the unconditional LGM. The estimates provided under the "Means" heading are fixed effects. For instance, the mean of the Intercept (INT) is 34.79, which reflects the estimate of the sample average level of depression (i.e., the model-implied estimate of DEP_1) at the first assessment (i.e., when the Slope, or Time, = 0). The Slope (SLP) mean reflects the estimate of how much the outcome changes on average with every unit increase in Time. Because this estimate is –3.62, this would be interpreted as indicating that, on average, depression scores decrease by 3.62 between each assessment period (i.e., this estimate

should be not be interpreted as reflecting the *total* reduction in depression across the entire active treatment phase). Importantly, the corresponding test ratio ($z = 3.62/0.41 = 8.91$, cf. Table 16.2) indicates there is a statistically significant reduction in depression scores for each unit increase in Time.

The estimates under the "Variances" heading are random effects (i.e., variability around the INT and SLP means). For instance, the SLP variance (13.31) is statistically significant ($p = .003$), indicating there is significant variability in the individual growth trajectories of depression during active treatment (i.e., significant individual differences in treatment change). The INT variance (246.99, $p < .001$) indicates considerable variability in patients' depression scores at the initial assessment. The "SLP WITH INT" estimate is the covariance of the two growth factors (–10.83). Although nonsignificant, the estimate reflects an inverse relationship between the Intercept and Slope (i.e., patients with higher scores at the initial assessment evidence steeper reductions in depression during active treatment).

Conditional LGM

After an acceptable unconditional LGM has been established (e.g., to verify the proper form of the growth trajectory), the researcher is usually interested in evaluating variables that may account for individual differences in change. A *conditional LGM* is an LGM where the growth factors are regressed onto background variables (i.e., covariates that may affect the growth trajectories). For example, a conditional LGM is depicted in the second path diagram in Figure 16.7, whereby the Intercept and Slope factors in the growth model of depression are regressed onto a Treatment dummy code (1 = CBT, 0 = nondirective treatment [ND]) with the intent of determining whether change in depression varies as a function of treatment assignment. The Mplus syntax and selected output (unstandardized solution) are presented in Table 16.3.

The conditional LGM fit the data well: $\chi^2(7) = 0.97$, $p = .99$ (not shown in Table 16.3). Although the unconditional LGM provided a good fit to the data, model fit should be reevaluated when other variables are added to the model. Of particular interest

Table 16.3 Mplus Syntax and Selected Output for the Conditional Latent Growth Model

```
TITLE: CONDITIONAL LATENT GROWTH MODEL OF DEPRESSION
DATA:
  FILE IS LGM.DAT;
  TYPE IS MEANS STD CORR;
  NOBS = 200;
VARIABLE:
  NAMES ARE DEP1 DEP2 DEP3 DEP4 TX;
  USEV = DEP1 DEP2 DEP3 DEP4 TX;
  ANALYSIS: ESTIMATOR=ML;
MODEL:    INT BY DEP1@1 DEP2@1 DEP3@1 DEP4@1;
            SLP BY DEP1@0 DEP2@1 DEP3@2 DEP4@3;
            [DEP1@0 DEP2@0 DEP3@0 DEP4@0 INT SLP];
            INT SLP ON TX;
OUTPUT: SAMPSTAT STANDARDIZED MODINDICES(4) TECH1;
```

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
INT	BY				
DEP1		1.000	0.000	999.000	999.000
DEP2		1.000	0.000	999.000	999.000
DEP3		1.000	0.000	999.000	999.000
DEP4		1.000	0.000	999.000	999.000
SLP	BY				

(*continued*)

Table 16.3 (Continued)

DEP1		0.000	0.000	999.000	999.000
DEP2		1.000	0.000	999.000	999.000
DEP3		2.000	0.000	999.000	999.000
DEP4		3.000	0.000	999.000	999.000
INT	ON				
TX		-2.052	2.193	-0.936	0.349
SLP	ON				
TX		-2.635	0.681	-3.870	0.000
SLP	WITH				
INT		-12.302	8.950	-1.374	0.169
Intercepts					
DEP1		0.000	0.000	999.000	999.000
DEP2		0.000	0.000	999.000	999.000
DEP3		0.000	0.000	999.000	999.000
DEP4		0.000	0.000	999.000	999.000
INT		35.814	1.672	21.421	0.000
SLP		-2.301	0.519	-4.432	0.000
Residual Variances					
DEP1		105.003	20.450	5.135	0.000
DEP2		105.890	14.081	7.520	0.000
DEP3		104.415	14.038	7.438	0.000
DEP4		95.009	19.933	4.766	0.000
INT		245.370	33.421	7.342	0.000
SLP		10.631	4.219	2.520	0.012

in this example are the estimates corresponding to the regression paths from the Treatment dummy code to the growth factors (i.e., INT ON TX, SLP ON TX). The unstandardized TX → INT path was –2.05, indicating that, on average, patients assigned to CBT scored 2.05 units less on DEP_1 (the initial depression assessment) than ND patients. However, this path was nonsignificant ($p = .35$), a desired result given treatment randomization. A significant TX → SLP path was obtained (–2.64, $p < .001$). Although the path diagram might suggest this is a direct effect (i.e., Treatment has a direct effect on the Slope), this path actually reflects a two-way interaction effect. Specifically, the effect of Time (i.e., the Slope) on the outcome variable varies as a function of treatment condition (a Time × Treatment interaction). The TX → SLP path estimate indicates how much more the Slope mean increases or decreases given a unit increase in TX. Because the sign of this estimate is negative, this indicates that CBT patients evidenced greater symptom reductions than ND patients. In other words, when TX increases by one unit (0 = ND, 1 = CBT), the relationship between Time and the depression decreases by 2.64 units (i.e., from each assessment wave to the next, the reduction in depression is 2.64 units larger for CBT patients than ND patients). In applied research studies, it would be most informative to plot the nature of this interaction effect in graphical format. The model-implied values of depression scores for the two treatment

conditions across the four assessments can be computed using the following equation (cf. Curran, Bauer, & Willoughby, 2004):

$$y|x = (\alpha_{INT} + \gamma_1 x) + (\alpha_{SLP} + \gamma_2 x)\lambda_T$$

where y = depression score at a given assessment;
x = treatment dummy code;
α_{INT} = Intercept mean;
α_{SLP} = Slope mean;
γ_1 = TX → INT path;
γ_2 = TX → SLP path; and
λ_T = value of Slope factor loading.

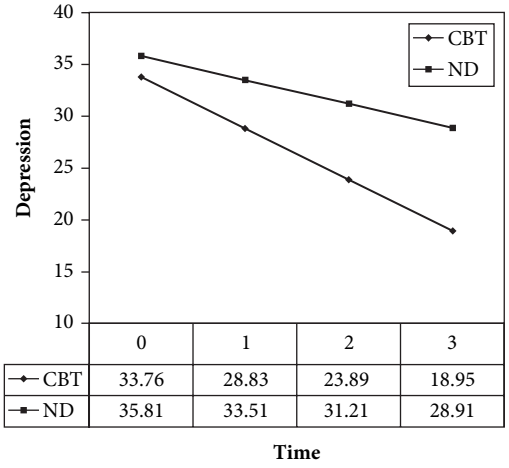
For instance, using the relevant estimates from Table 16.3, the model-implied estimate of the depression score at the fourth assessment ($\lambda_T = 3$) for CBT patients ($x = 1$) is:

$$DEP_4 = [35.81 + -2.05(1)] + [-2.30 + -2.63(1)]3 = 18.95$$

As seen in Figure 16.8, the other seven depression score means are calculated using this equation and plotted in an X/Y graph to illustrate the model-implied treatment response trajectories of CBT and ND patients (for applied examples, see Brown, 2007; Brown & Rosellini, 2011).

Applications and Extensions of LGM

Although LGM has yet to be widely adopted in treatment outcome trials, this analytic framework is well suited to this research domain. LGM requires a minimum of three repeated observations (and more than three is preferred), so treatment outcome studies should be designed accordingly to make use of this method (i.e., multiple assessments during both the active treatment and treatment follow-up phases). As noted earlier, LGM has several advantages over repeated measures analysis of variance (ANOVA, perhaps the most common analytic approach to treatment outcome data), including greater statistical power and the ability to formally address the shape of growth and a measurement error theory. Unlike ANOVA, the sphericity assumption is not germane to LGM because the focus is on individual growth trajectories rather than mean change between adjacent time points. Multilevel modeling (MLM) is another method for analyzing time-series data (e.g., using the HLM program, or mixed regression routines in SAS and SPSS/PASW). Whereas LGM can be specified to render the same results as MLM under some conditions, LGM offers considerably more modeling flexibility, including the ability to estimate regressions among random effects (i.e., directional paths between growth factors), to embed the latent trajectory model within a larger SEM model (e.g., allow the LGM growth factors to predict a distal latent variable outcome), to simultaneously estimate growth models in multiple groups (e.g., separate models for treatment and treatment comparison groups; cf. Muthén & Curran, 1997), and to use latent variables as repeated outcomes instead of single indicators (discussed later).



	0	1	2	3
CBT	33.76	28.83	23.89	18.95
ND	35.81	33.51	31.21	28.91

Figure 16.8 Model-implied trajectories of depression scores during active treatment phase (CBT = cognitive-behavioral treatment, ND = nondirective treatment).

Indeed, LGM should be considered an analytic method of choice for any longitudinal study where systematic change in one or more outcomes is anticipated (e.g., laboratory studies involving the process of habituation, developmental studies of reading or other emerging academic abilities in children). However, LGM is not well suited to time-series data where systematic change is not evident (e.g., as perhaps would be the case in studies of clinical processes in community or other nonclinical samples). The AR model may be more appropriate in this situation where the focus is on residualized change (i.e., ability of a predictor to explain unique variance in the temporal outcome after controlling for the autoregressive effect) instead of individual differences in continuous developmental trajectories over time. The remaining portion of this section will present some of the many possible extensions of the LGM.

Nonlinear growth. The examples discussed in this chapter thus far have involved linear change; that is, disregarding sampling error, the amount of improvement in depression during the active treatment phase is the same for the various time-adjacent assessment

points. There are many scenarios, however, where linear growth is untenable. For instance, in a laboratory study of habituation, the dependent variable may be expected to level off after a period of steady decline. There are several ways nonlinear growth can be modeled in LGM. Three of these approaches are depicted as path diagrams in Figure 16.9: LGM with freely estimated time scores, the polynomial growth model, and the piecewise growth model.

As shown in Figure 16.9, the **LGM with freely estimated time scores** entails the same two growth factors as the linear growth LGM (i.e., Intercept and Slope). However, unlike the linear LGM, only two of the Slope factor loadings are fixed to predetermined values. In virtually all instances, one of these loadings is fixed to zero (the centering point) and the other is fixed to 1.0. The remaining Slope factor loadings are freely estimated. In the example in Figure 16.9, the Slope loadings corresponding to the middle two time points are freely estimated. Thus, instead of providing a formal test of anticipated change (i.e., evaluation of the goodness of fit

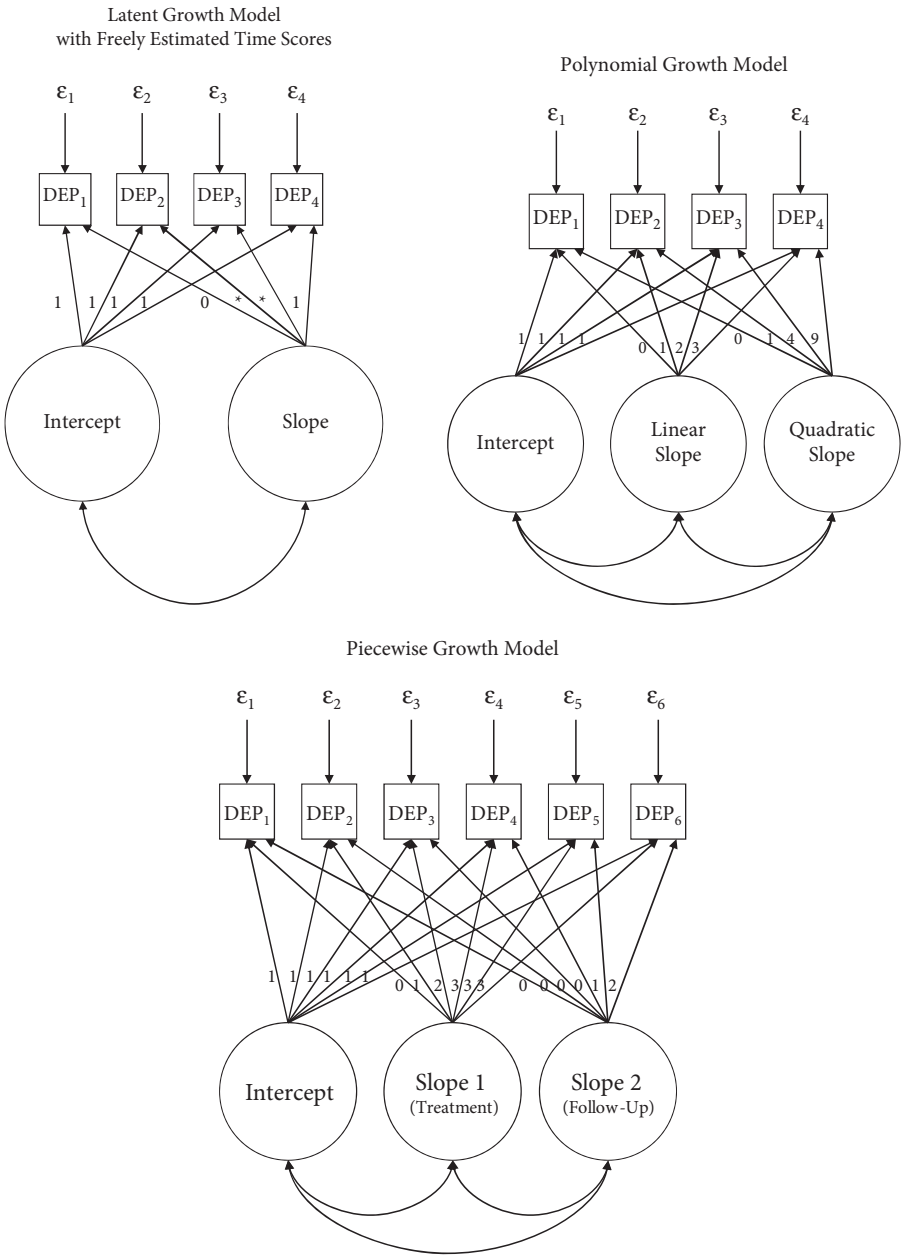


Figure 16.9 Nonlinear growth models. DEP = depression; * = freely estimated parameter.

of the hypothesized shape of growth), the resulting shape of the growth trajectory is determined by the sample data. Although the atheoretical nature of this specification could be viewed as a drawback, the LGM with freely estimated time scores may be appropriate in situations where the shape of growth has little substantive value, or when the number of repeated observations is limited (thereby precluding statistical identification of more elaborate models, such as the piecewise growth model). Unlike the linear LGM, the Slope mean cannot be interpreted as the amount of constant change across all time points. Rather, the Slope mean is the estimated change in the outcome for a unit increase in Time. The interpretation of the Slope mean in this nonlinear LGM depends on what Slope factor loadings were fixed to values of zero and 1.0. In the Figure 16.9 example, the first and fourth factor loadings were fixed to zero and 1.0, respectively. Thus, the Intercept is centered on the first observation, and the Slope mean will reflect the estimated amount of change in the outcome from the first to the fourth assessment. The fixed Slope factor loadings should be selected so that a one-unit change corresponds to the time points of substantive interest. For instance, the first model in Figure 16.9 could be respecified with Slope factor loadings of 0 1 ** (* = freely estimated) if the researcher was interested in interpreting the Slope with respect to change occurring between the first and second observation. Applied clinical research examples of this approach can be found in Brown (2007) and Brown and Rosellini (2011).

Another method of modeling nonlinear change is the **LGM with polynomial growth factors**. Unlike the LGM with free time scores, the anticipated nonlinear shape of change is prespecified in this model. The second path diagram in Figure 16.9 is a quadratic growth model. This model is appropriate when a single bend is expected in the growth trajectory (e.g., as might be the case in the habituation example mentioned earlier). This approach involves three growth factors: the Intercept, the Linear Slope, and the Quadratic Slope. As seen in Figure 16.9, the Intercept and Linear Slope are specified in the same manner as the linear LGM. The Quadratic Slope factor loadings are fixed to values that equal the Linear Slope loadings raised to the second power (using the same logic as in the formation of power polynomials in nonlinear multiple regression; cf. Cohen, Cohen, West, & Aiken, 2003). The Quadratic Slope mean provides the estimate of how much the magnitude of change is accelerating or decelerating over time. For instance, if statistically significant (and negative-signed) Linear and Quadratic Slope means were obtained in the treatment outcome example, this would indicate that depression scores were decreasing over time and the magnitude of this symptom reduction was increasing as treatment progressed. Although very rare in applied research, the quadratic model can be readily extended to further polynomial functions (e.g., cubic: two bends in the growth curve) if there is a sufficient number of repeated assessments and if substantively justified. Examples of quadratic LGMs can be found in Harrison, Blozois, and Stuifbergen (2008), Orth, Trzesniewski, and Robins (2010), and Vazonyi and Keiley (2007). For instance, using a cohort-sequential design, Harrison and colleagues (2008) found that self-esteem has a quadratic trajectory across the adult lifespan, increasing during young and middle adulthood, peaking at about age 60, and declining in old age.

A drawback of polynomial LGMs is that they can be difficult to interpret, particularly if background variables are included in an attempt to explain variance in the growth factors (e.g., TX → Quadratic Slope). Moreover, the polynomial growth factors often have considerably less variance than the Intercept and Linear Slope factors, thereby reducing the statistical power of predictors to account for nonlinear change. **Piecewise growth models** are an alternative method for estimating nonlinear trajectories. As with polynomial LGM, nonlinear change is estimated in the piecewise model through the use of additional growth factors. However, unlike the polynomial LGM, the additional growth factors are prespecified by the researchers in accord with their theory about transitional points in the developmental trajectory. Alternatively, these models can be specified to focus on estimating change during a specific time period of substantive interest within a broader longitudinal design. Although underutilized in this area thus far, piecewise models are very well suited to treatment outcome research where it is usually expected that symptom reduction will be most pronounced during the active treatment phase, and continued but less extensive improvement will be evident during the follow-up phase. A piecewise LGM of this nature is presented in Figure 16.9. In this model, one Linear Slope factor is specified to estimate change during active treatment (DEP_1 through DEP_4), and a second Linear Slope factor estimates change during follow-up (DEP_4 through DEP_6). These models can also be specified with more than one Intercept factor, if substantively useful (e.g., another Intercept at the transition point in the developmental process). Another advantage of this approach is that each "piece" in the model can have its own predictors, which is useful for testing

the hypothesis that a covariate will predict change during one period of time but not another (e.g., separate predictors of active treatment response and maintenance of treatment gains). The covariance between the two slopes can also be of substantive interest. For instance, a significant covariance between the Slope factors in Figure 16.9 might imply that, on average, patients who evidenced the most gains during active treatment tended to improve the most during follow-up. Although offering considerable analytic potential, piecewise LGMs are rare in the applied literature. A study by Clark and colleagues (2005) was one of the first applications of this approach to treatment outcome data (i.e., separate growth factors for active treatment and follow-up in an open-label trial of fluoxetine for childhood anxiety disorders). A paper by Llabre, Spitzer, Saab, and Scheiderman (2001) demonstrates how this method can be put to good use in the analysis of laboratory challenge data (e.g., separate growth factors for systolic blood pressure reactivity and recovery from the cold pressor test). Studies by Jones and Meredith (2000) and Crawford, Pentz, Chou, Li, and Dwyer (2003) illustrate applications of this approach for modeling developmental processes in longitudinal survey samples (e.g., psychological health, substance use). The reader is referred to Flora (2008) for further information on the specification and interpretation of piecewise LGMs.

Multivariate models. LGM can be extended to incorporate repeated measures from more than one dependent variable. Three examples of such models are presented in Figure 16.10: the parallel process LGM, the LGM with time-varying covariates, and the LGM with latent variable outcomes. In the **parallel process LGM**, the growth trajectories of two or more dependent measures are modeled simultaneously. This allows for the evaluation of the temporal relationship between two or more constructs at the level of the random growth factors. For instance, in the first path diagram in Figure 16.10, separate latent curve processes are modeled for depression and neuroticism. In the specification shown, the growth factors of both processes are simply allowed to covary. The covariance of the two Slope factors reflects the correspondence between the rates of change in these two constructs over time (i.e., does change in depression covary with change in neuroticism). Given how the growth processes in Figure 16.10 are centered, the covariance of the Intercepts represents the relationship between depression and neuroticism at the first assessment. It is also possible to specify regression paths among the growth factors. For example, it might be of interest to examine whether the rate of change in depression is predicted by the initial status of neuroticism (holding initial status of depression constant). Parallel process LGMs of this nature were reported in a study by Brown (2007) whereby neuroticism predicted the temporal course of the DSM-IV disorder constructs of generalized anxiety disorder and social phobia, but not vice versa (i.e., the initial levels of the disorder constructs did not predict the course of neuroticism). Other applications of the parallel process LGMs in the clinical literature include studies by Curran, Stice, and Chassin (1997) and Crawford and colleagues (2003). Researchers have extended the parallel process LGM to estimate time-specific relations among the repeated measures in addition to the underlying random trajectories. Prominent versions of these hybrid models include the latent difference score model (McArdle, 2001, 2009) and the autoregressive latent trajectory model (Bollen & Curran, 2004).

The second path diagram in Figure 16.10 is an **LGM with time-varying covariates** (TVC). Unlike the parallel process LGM, although repeated measures of a second construct are included (life stress), a growth process for this construct is not estimated. This type of model hypothesizes that the second construct (i.e., the TVC) does not evidence systematic change, but it plays an important role in accounting for variance in the repeated measures of the first construct, over and beyond the latter's underlying trajectory process. An assumption of the LGMs presented earlier (cf. Fig. 16.7) is that the repeated measures of the outcome variable are determined entirely by the underlying growth factors. Alternatively, an LGM with TVCs hypothesizes that the repeated measures of a construct are determined in part by the underlying growth process but also by other time-specific (or time-lagged) constructs. For example, the model in Figure 16.10 predicts that the time-specific measures of depression would be related to time-specific measures of life stress (the TVC) over and beyond the trajectory process underlying depression. However, a separate growth trajectory for life stress was not modeled based on the premise that this variable would not change systematically over the follow-up period. Applied examples of LGMs with TVCs include studies by Harrison and colleagues (2008) and Hussong and colleagues (2004). Harrison and colleagues (2008), for instance, used this approach to examine whether functional limitations due to multiple sclerosis were predicted by time-specific measures of social support, over and beyond the latent trajectory process of functional limitations.

The final path diagram in Figure 16.10 is an **LGM where the repeated measures are represented**

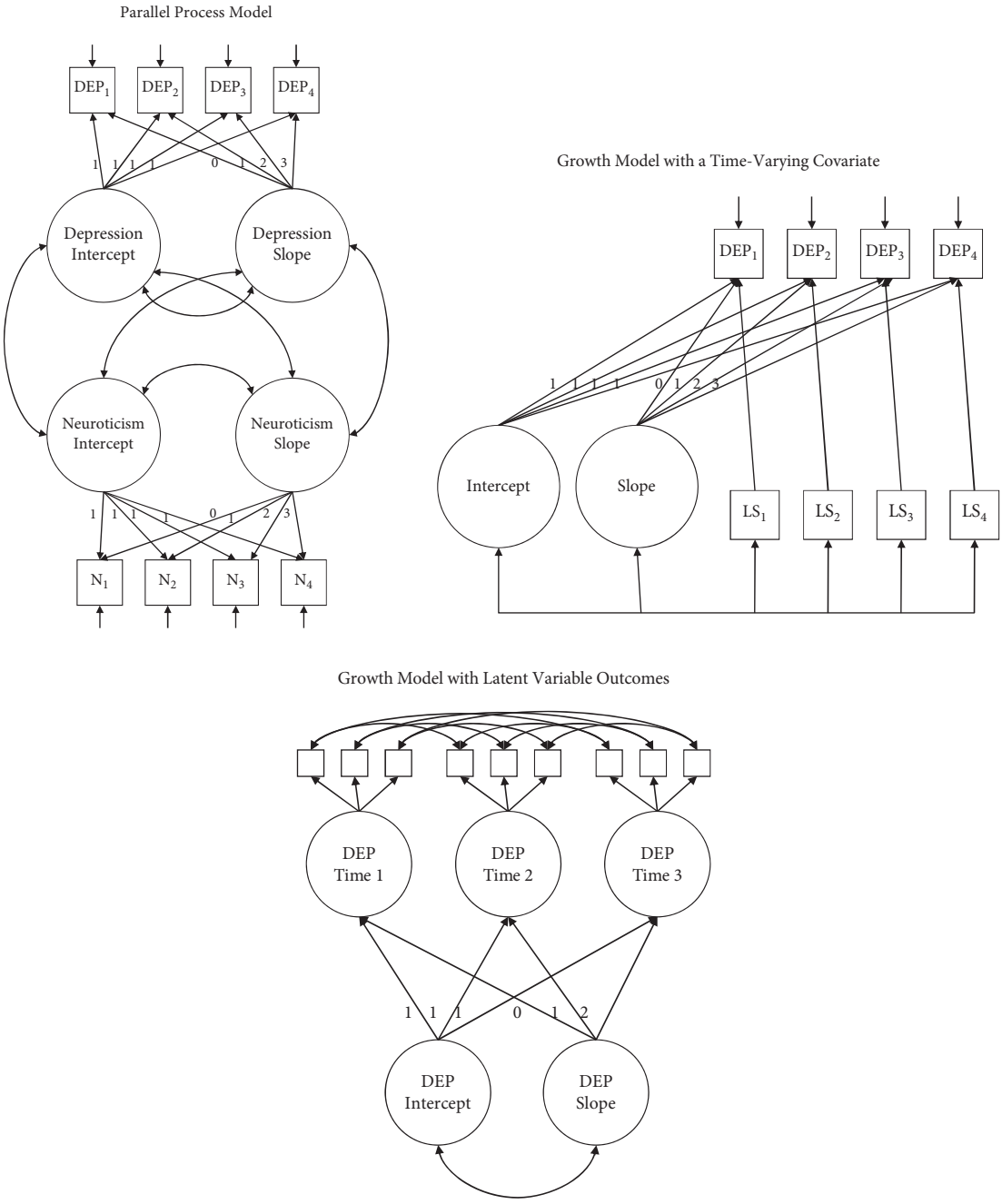


Figure 16.10. Multivariate growth models. DEP = depression, N = neuroticism, LS = life stress.

by latent variables rather than single indicators. Because it resides in the SEM analytic framework, LGM can draw on the key strengths of SEM. One such advantage is the ability to model the trajectory of latent variables defined by multiple indicators. In the model shown in Figure 16.10, three different repeated measures of depression (various questionnaire and clinical ratings) were obtained and these measures served as the indicators for the latent variable of this construct at each time point. Accordingly, growth is analyzed within a latent variable measurement model such that the depression outcomes are theoretically free of measurement error and are estimated in the context of a formal evaluation of longitudinal measurement invariance (e.g., time-specific variance and measurement error variance are not confounded; statistical power is fostered by smaller standard errors of the growth factor parameters). As seen in Figure

16.10, another advantage of this model is that the residual variances for the same indicator over time can be permitted to correlate (e.g., removal of variance due to testing effects from the growth factor parameters). Applications of this approach are very rare in the clinical literature, although studies by Beauchaine, Webster-Stratton, and Reid (2005), Brown (2007), and Brown and Rosellini 2011) are exceptions.

Summary

This introductory overview of SEM discussed its applications to the study of psychopathology. In addition, we provided examples and syntax for some common types of models, described the multistep process of designing and evaluating models, and reviewed some of the advantages of using SEM over single-indicator–based analyses. Researchers are encouraged to consider SEM in their analytic design and to take a "hands-on" approach to learning the details of its execution. Conducting SEM is more straightforward than many might think, and the methods are well suited to answering important questions about construct validity and the classification, etiology, course, and treatment of psychopathology.

References

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Baschnagel, J. S., O'Connor, R. M., Colder, C. R., & Hawk, L. W., Jr. (2005). Factor structure of posttraumatic stress among western New York undergraduates following the September 11th terrorist attack on the World Trade Center. *Journal of Traumatic Stress*, *18*, 677–684.

Beauchaine, T. P., Webster-Stratton, C., & Reid, M. J. (2005). Mediators, moderators, and predictors of 1-year outcomes among children treated for early-onset conduct problems: A latent growth curve analysis. *Journal of Consulting and Clinical Psychology*, *73*, 371–388.

Beauducel, A., & Wittman, W. W. (2005). Simulation study on fit indices in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*, 41–75.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research*, *32*, 336–383.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Brown, T. A. (2007). Temporal course and structural relationships among dimensions of temperament and DSM-IV anxiety and mood disorder constructs. *Journal of Abnormal Psychology*, *116*, 313–328.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology*, *107*, 179–192.

Brown, T. A., & Rosellini, A. J. (2011). The direct and interactive effects of neuroticism and life stress on the severity and longitudinal course of depressive symptoms. *Journal of Abnormal Psychology*, *120*, 844–856.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campbell-Sills, L. A., Liverant, G., & Brown, T. A. (2004). Psychometric evaluation of the Behavioral Inhibition/Behavioral Activation Scales (BIS/BAS) in large clinical samples. *Psychological Assessment*, *16*, 244–254.

Clark, D. B., Birmaher, B., Axelson, D., Monk, K., Kalas, C., Ehmann, M., et al. (2005). Fluoxetine for the treatment of childhood anxiety disorders: Open-label, long-term extension to a controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 1263–1270.

Cohen, J., Cohen, P., West, S. G., & AIken, L. S. (2003). *Applied multiple/regression correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Crawford, A. M., Pentz, M. A., Chou, C. P., Li, C., & Dwyer, J. H. (2003). Parallel developmental trajectories of sensation seeking and regular substance use in adolescents. *Psychology of Addictive Behaviors*, *17*, 179–192.

Curran, P. J., Bauer, D. J., & Willoughby, M. T. (2004). Testing main effects and interactions in latent curve analysis. *Psychological Methods*, *9*, 220–237.

Curran, P. J., Stice, E., & Chassin, L. (1997). The relation between adolescent alcohol use and peer alcohol use: A longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology*, *65*, 130–140.

Elklit, A., & Shevlin, M. (2007). The structure of PTSD symptoms: A test of alternative models using confirmatory factor analysis. *British Journal of Clinical Psychology*, *46*, 299–313.

Farrell, A. D. (1994). Structural equation modeling with longitudinal data: Strategies for examining group differences and reciprocal relationships. *Journal of Consulting and Clinical Psychology*, *62*, 477–487.

Flora, D. B. (2008). Specifying piecewise latent trajectory models for longitudinal data. *Structural Equation Modeling*, *15*, 513–533.

Goldberg, D. P., Krueger, R. F., & Andrews, G. (2009). Emotional disorders: Cluster 4 of the proposed meta-structure for *DSM-V* and ICD-11. *Psychological Medicine*, *39*, 2043–2059.

Harrison, T., Blozois, S., & Stuifbergen, A. (2008). Longitudinal predictors of attitudes toward aging among women with multiple sclerosis. *Psychology and Aging*, *23*, 823–832.

Hettema, J. M., An, S. S., Bukszar, J., van den Oord, E. J., Neale, M. C., Kendler, K. S., et al. (2008). Catechol-O-methyltransferase contributes to genetic susceptibility shared among anxiety spectrum phenotypes. *Biological Psychiatry*, *64*, 302–310.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hussong, A. M., Curran, P. J., Moffitt, T. E., Caspi, A., & Carrig, M. M. (2004). Substance abuse hinders desistance in young adults' antisocial behavior. *Development and Psychopathology*, *16*, 1029–1046.

Jones, C. L., & Meredith, W. (2000). Developmental paths of psychological health from early adolescence to later adulthood. *Psychology and Aging*, *15*, 351–360.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165–172.

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of

12-month DSM–IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *62*, 617–627.

King, D. W., Leskin, G. A., King, L. A., & Weathers, F. W. (1998). Confirmatory factor analysis of the clinician-administered PTSD Scale: Evidence for the dimensionality of posttraumatic stress disorder. *Psychological Assessment*, *10*, 90–96.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kollman, D. M., Brown, T. A., & Barlow, D. H. (2009). The construct validity of acceptance: A multitrait-multimethod investigation. *Behavior Therapy*, *40*, 205–218.

Kramer, M. D., Krueger, R. F., & Hicks, B. M. (2008). The role of internalizing and externalizing liability factors in accounting for gender differences in the prevalence of common psychopathological syndromes. *Psychological Medicine*, *38*, 51–61.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*, 921–926.

Krueger, R. F., Caspi, A., Moffitt, T. E., & Silva, P. A. (1998). The structure and stability of common mental disorders (*DSM–III–R*): A longitudinal-epidemiological study. *Journal of Abnormal Psychology*, *107*, 216–227.

Krueger, R. F., Chentsova-Dutton, Y. E., Markon, K. E., Goldberg, D., & Ormel, J. (2003). A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *Journal of Abnormal Psychology*, *112*, 437–447.

Krueger, R. F., & South, S. C. (2009). Externalizing disorders: Cluster 5 of the proposed meta-structure for *DSM-V* and ICD-11. *Psychological Medicine*, *39*, 2061–2070.

Llabre, M. M., Spitzer, S. B., Saab, P. G., & Scheiderman, N. (2001). Piecewise latent growth curve modeling of systolic blood pressure reactivity and recovery from the cold pressor test. *Psychophysiology*, *38*, 951–960.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Psychology Press.

Markon, K. E. (2010). Modeling psychopathology structure: A symptom-level analysis of Axis I and II disorders. *Psychological Medicine*, *40*, 273–288.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D.Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Miller, M. W., Fogler, J. F., Wolf, E. J., Kaloupek, D. G., & Keane, T. M. (2008). The internalizing and externalizing structure of psychiatric comorbidity in combat veterans. *Journal of Traumatic Stress*, *21*, 58–65.

Miller, M. W., Vogt, D. S., Mozley, S. L., Kaloupek, D. G., & Keane, T. M. (2006). PTSD and substance-related problems: The mediating roles of disconstraint and negative emotionality. *Journal of Abnormal Psychology*, *115*, 369–379.

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology*, *49*, 377–412.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *4*, 599–620.

Nock, N. L., Wang, X., Thompson, C. L., Song, Y., Baechle, D., Raska, P., et al. (2009). Defining genetic determinants of the metabolic syndrome in the Framingham Heart Study using association and structural equation modeling methods. *BMC Proceedings*, *3*, S50.

Orth, U., Trzesniewski, K. H., & Robins, R. W. (2010). Self-esteem development from young adulthood to old age: A cohort-sequential longitudinal study. *Journal of Personality and Social Psychology*, *98*, 645–658.

Palmieri, P. A., Weathers, F. W., Difede, J., & King, D. W. (2007). Confirmatory factor analysis of the PTSD Checklist and the Clinician-Administered PTSD Scale in disaster workers exposed to the World Trade Center ground zero. *Journal of Abnormal Psychology*, *116*, 329–341.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Simms, L. J., Watson, D., & Doebbeling, B. N. (2002). Confirmatory factor analyses of posttraumatic stress symptoms in deployed and nondeployed veterans of the Gulf War. *Journal of Abnormal Psychology*, *111*, 637–647.

Slade, T., & Watson, D. (2006). The structure of common DSM–IV and ICD-10 mental disorders in the Australian general population. *Psychological Medicine*, *36*, 1593–1600.

Vazonyi, A. T., & Keiley, M. K. (2007). Normative developmental trajectories of aggressive behaviors in African American, American Indian, Asian American, Caucasian, and Hispanic children and early adolescents. *Journal of Abnormal Child Psychology*, *35*, 1047–1062.

Vollebergh, W. A., Iedema, J., Bijl, R. V., de Graaf, R., Smit, F., & Ormel, J. (2001). The structure and stability of common mental disorders. *Archives of General Psychiatry*, *58*, 597–603.

Wolf, E. J., Miller, M. W., Krueger, R. F., Lyons, M. J., Tsuang, M. T., & Koenen, K. C. (2010). PTSD and the genetic structure of comorbidity. *Journal of Abnormal Psychology*, *119*, 320–330.

Wothke, W. A. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.

Yeh, P.-H., Zhu, H., Nicoletti, M. A., Hatch, J. P., Brambilla, P., & Soares, J. C. (2010). Structural equation modeling and principal component analysis of gray matter volumes in major depressive and bipolar disorders: Differences in latent volumetric structure. *Psychiatry Research: Neuroimaging*, *184*, 177–185.

CHAPTER

17 Meta-analysis in Clinical Psychology Research

Andy P. Field

Abstract

Meta-analysis is now the method of choice for assimilating research investigating the same question. This chapter is a nontechnical overview of the process of conducting meta-analysis in the context of clinical psychology. We begin with an overview of what meta-analysis aims to achieve. The process of conducting a meta-analysis is then described in six stages: (1) how to do a literature search; (2) how to decide which studies to include in the analysis (inclusion criteria); (3) how to calculate effect sizes for each study; (4) running a basic meta-analysis using the metaphor package for the free software R; (5) how to look for publication bias and moderators of effect sizes; and (6) how to write up the results for publication.

Key Words: Meta-analysis, effect sizes, publication bias, moderator analysis, R

Introduction

Meta-analysis has become an increasingly popular research methodology, with an exponential increase in papers published seen across both social sciences and science in general. Field (2009) reports data showing that up until 1990 there were very few studies published on the topic of meta-analysis, but after this date the use of this tool has been on a meteoric increase. This trend has occurred in clinical psychology too. Figure 17.1 shows the number of articles with "meta-analysis" in the title published within the domain of "clinical psychology" since the term "meta-analysis" came into common usage. The data show a clear increase in publications after the 1990s, and a staggering acceleration in the number of published meta-analyses in this area in the past 3 to 5 years. Meta-analysis has been used to draw conclusions about the causes (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van Ijzendoorn, 2007; Brewin, Kleiner, Vasterling, & Field, 2007; Burt, 2009; Chan, Xu, Heinrichs, Yu, & Wang, 2010; Kashdan, 2007; Ruocco, 2005), diagnosis (Bloch, Landeros-Weisenberger, Rosario, Pittenger, & Leckman, 2008; Cuijpers, Li, Hofmann, & Andersson, 2010), and preferred treatments (Barbato & D'Avanzo, 2008; Bradley, Greene, Russ, Dutra, & Westen, 2005; Cartwright-Hatton, Roberts, Chitsabesan, Fothergill, & Harrington, 2004; Covin, Ouimet, Seeds, & Dozois, 2008; Hendriks, Voshaar, Keijsers, Hoogduin, & van Balkom, 2008; Kleinstaeuber, Witthoeft, & Hiller, 2011; Malouff, Thorsteinsson, Rooke, Bhullar, & Schutte, 2008; Parsons & Rizzo, 2008; Roberts, Kitchiner, Kenardy, Bisson, & Psych, 2009; Rosa-Alcazar, Sanchez-Meca, Gomez-Conesa, & Marin-Martinez, 2008; Singh, Singh, Kar, & Chan, 2010; Spreckley & Boyd, 2009; Stewart & Chambless, 2009; Villeneuve, Potvin, Lesage, & Nicole, 2010) of a variety of mental health problems. This illustrative selection of articles shows that meta-analysis has been used to determine the efficacy of behavioral, cognitive, couple-based, cognitive-behavioral (CBT), virtual reality, and psychopharmacological interventions and on problems as diverse as

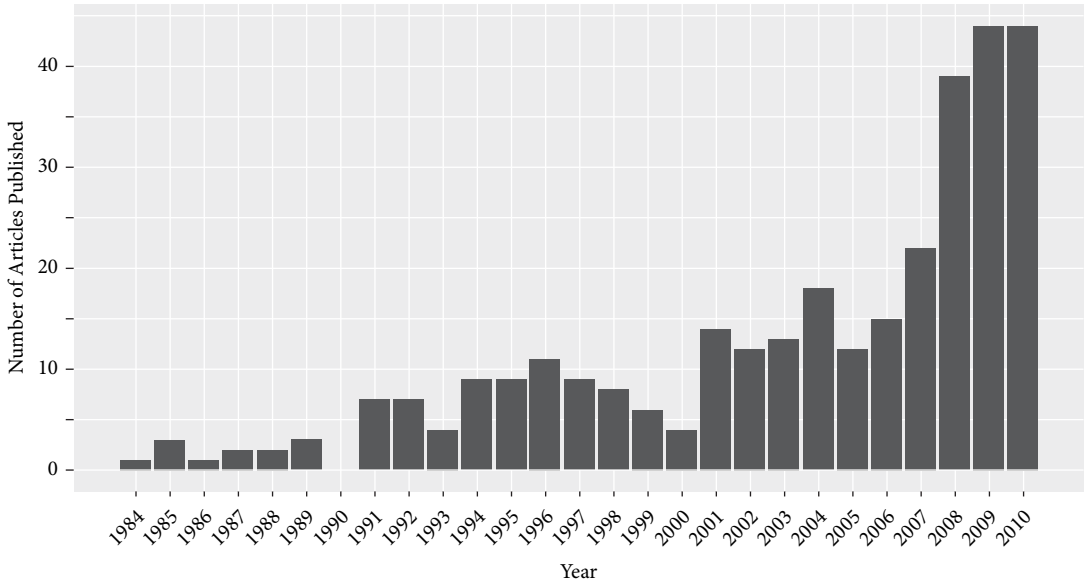


Figure 17.1 The number of studies using meta-analysis in clinical psychology.

schizophrenia, anxiety disorders, depression, chronic fatigue, personality disorders, and autism. There is little in the world of clinical psychology that has not been subjected to meta-analysis.

This chapter provides a practical introduction to meta-analysis. For mathematical details, see other sources (e.g., H. M. Cooper, 2010; Field, 2001, 2005a, 2009; Field & Gillett, 2010; Hedges & Olkin, 1985; Hedges & Vevea, 1998; Hunter & Schmidt, 2004; Overton, 1998; Rosenthal & DiMatteo, 2001; Schulze, 2004). This chapter overviews the important issues when conducting meta-analysis and shows an example of how to conduct a meta-analysis using the free software R (R Development Core Team, 2010).

What Is Meta-analysis?

Clinical psychologists are typically interested in reaching conclusions that can be applied generally. These questions might whether CBT is efficacious as a treatment for obsessive-compulsive disorder (Rosa-Alcazar et al., 2008), whether antidepressant medication treats the negative symptoms of schizophrenia (Singh et al., 2010), whether virtual reality can be effective in treating specific phobias (Parsons & Rizzo, 2008), whether school-based prevention programs reduce anxiety and/or depression in youth (Mychailyszyn, Brodman, Read, & Kendall, 2011), whether there are memory deficits for emotional information in posttraumatic stress disorder (PTSD; Brewin et al., 2007), what the magnitude of association between exposure to disasters and youth PTSD is (Furr, Corner, Edmunds, & Kendall, 2010), or what the magnitude of threat-related attentional biases in anxious individuals is (Bar-Haim, et al., 2007). Although answers to these questions may be attainable in a single study, single studies have two limitations: (1) they are at the mercy of their sample size because estimates of effects in small samples will be more biased than large sample studies and (2) replication is an important means to deal with the problems created by measurement error in research (Fisher, 1935). Meta-analysis pools the results from similar studies in the hope of generating more accurate estimates of the true effect in the population. A meta-analysis can tell us:

1. *The mean and variance of underlying population effects*—for example, the effects in the population of conducting CBT with depressed adolescents compared to waitlist controls. You can also compute confidence intervals for the population effects.
2. *Variability in effects across studies.* It is possible to estimate the variability between effect sizes across studies (the homogeneity of effect sizes). There is accumulating evidence that effect sizes should be heterogeneous across studies (see, e.g., National Research Council, 1992). Therefore, variability statistics should be reported routinely. (You will often see significance tests reported for these estimates of variability; however, these tests typically have low power and are probably best ignored.)
3. *Moderator variables.* If there is variability in effect sizes, and in most cases there is

(Field, 2005a), this variability can be explored in terms of moderator variables (Field, 2003b; Overton, 1998). For example, we might find that attentional biases to threat in anxious individuals are stronger when picture stimuli are used to measure these biases than when words are used.

A Bit of History

More than 70 years ago, Fisher and Pearson discussed ways to combine studies to find an overall probability (Fisher, 1938; Pearson, 1938), and over 60 years ago, Stouffer presented a method for combining effect sizes (Stouffer, 1949). The roots of meta-analysis are buried deep within the psychological and statistical earth. However, clinical psychology has some claim over the popularization of the method: in 1977, Smith and Glass published an influential paper in which they combined effects from 375 studies that had looked at the effects of psychotherapy (Smith & Glass, 1977). They concluded that psychotherapy was effective, and that the type of psychotherapy did not matter. A year earlier, Glass (1976) published a paper in which he coined the term "meta-analysis" (if this wasn't the first usage of the term, then it was certainly one of the first) and summarized the basic principles. Shortly after these two seminal papers, Rosenthal published an influential theoretical paper on meta-analysis, and a meta-analysis combining 345 studies to show that interpersonal expectancies affected behavior (Rosenthal, 1978; Rosenthal & Rubin, 1978). It is probably fair to say that these papers put "meta-analysis" in the spotlight of psychology. However, it was not until the early 1980s that three books were published by Rosenthal (1984, 1991), Hedges and Olkin (1985), and Hunter and Schmidt (1990). These books were the first to provide detailed and accessible accounts of how to conduct a meta-analysis. Given a few years for researchers to assimilate these works, it is no surprise that the use and discussion of meta-analysis accelerated after 1990 (see Fig. 17.1). The even more dramatic acceleration in the number of published meta-analyses in the past 5 years is almost certainly due to the widespread availability of computer software packages that make the job of meta-analysis easier than before.

Computer Software for Doing Meta-analysis

An Overview of the Options

There are several standalone packages for conducting meta-analyses: for example, the Cochrane Collaboration's Review Manager (RevMan) software (The Cochrane Collaboration, 2008). There is also a package called Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2005). There are two add-ins for Microsoft Excel: Mix (Bax, Yu, Ikeda, Tsuruta, & Moons, 2006) and MetaEasy (Kontopantelis & Reeves, 2009). These packages implement many different meta-analysis methods, convert effect sizes, and create plots of study effects. Although it is not 100 percent clear from their website, Comprehensive Meta-Analysis appears to be available only for Windows, Mix works only with Excel 2007 and 2010 in Windows, and MetaEasy works with Excel 2007 (again Windows). RevMan uses Java and so is available for Windows, Linux, and MacOS operating systems. Although RevMan and MetaEasy are free and Mix comes in a free light version, Comprehensive Meta-Analysis and the pro version of Mix are commercial products.

SPSS (a commercial statistics package commonly used by clinical psychologists) does not incorporate a menu-driven option for conducting meta-analysis, but it is possible to use its syntax to run a meta-analysis. Field and Gillett (2010) provide a tutorial on meta-analysis and also include syntax files and examples showing how to run a meta-analysis using SPSS. Other SPSS syntax files can be obtained from Lavesque (2001) and Wilson (2004).

Meta-analysis can also be conducted with *R* (R Development Core Team, 2010), a freely available package for conducting a staggering array of statistical procedures. *R* is free, open-source software available for Windows, MacOS, and Linux that is growing in popularity in the psychology community. Scripts for running a variety of meta-analysis procedures on *d* are available in the *meta* package that can be installed into *R* (Schwarzer, 2005). However, my favorite package for conducting meta-analysis in *R* is *metafor* (Viechtbauer, 2010) because it has functions to compute effect sizes from raw data, can work with a wide array of different effect sizes (*d, r*, odds ratios, relative risks, risk differences, proportions, and incidence rates), produces publication-standard graphics, and implements moderator analysis and fixed- and random-effects methods (more on this later). It is a brilliant package, and given that it can be used for free across Windows, Linux, and MacOS, I have based this chapter on using this package within *R*.

Getting Started with R

R (R Development Core Team, 2010) is an environment/language for statistical analysis and is the

fastest-growing statistics software. *R* is a command language: we type in commands that we then execute to see the results. Clinical psychologists are likely to be familiar with the point-and-click graphical user interfaces (GUIs) of packages like SPSS and Excel, and so at first *R* might appear bewildering. However, I will walk through the process step by step assuming that the reader has no knowledge of *R*. I cannot, obviously, explain everything there is to know about *R*, and readers are advised to become familiar with the software by reading one of the many good introductory books (e.g., Crawley, 2007; Quick, 2010; Verzani, 2004; Zuur, Ieno, & Meesters, 2009), the best of which, in my entirely objective opinion, is Field, Miles, and Field (2012).

Once you have installed *R* on your computer and opened the software, you will see the console window, which contains a prompt at which you type commands. Once a command has been typed, you press the return key to execute it. You can also write commands in a script window and execute them from there, which is my preference—see Chapter 3 of Field and colleagues (2012). *R* comes with a basic set of functionality, which can be expanded by installing packages stored at a central online location that has mirrors around the globe. To install the *metafor* package you need to execute this command:

```
install.packages("metafor")
```

This command installs the package (you need to be connected to the Internet for this command to work). You need to install the package only once (although whenever you update *R* you will have to reinstall it), but you need to load the package every time that you want to use it. You do this by executing this command:

```
library(metafor)
```

The *library()* function tells *R* that you want to use a package (in this case *metafor*) in the current session. If you close the program and restart it, then you would need to re-execute the library command to use the *metafor* package.

The Six Basic Steps of Meta-analysis: An Example

Broadly speaking, there are six sequential steps to conducting a quality meta-analysis: (1) Do a literature search; (2) Decide on inclusion criteria; (3) Calculate the effect sizes; (4) Do the basic meta-analysis; (5) Do some more advanced analysis; and (6) Write it up. In this chapter, to illustrate these six steps, we will use a real dataset from a meta-analysis in which I was involved (Hanrahan, Field, Jones, & Davey, 2013). This meta-analysis looked at the efficacy of cognitive-based treatments for worry in generalized anxiety disorder (GAD), and the part of the analysis that we will use here simply aimed to estimate the efficacy of treatment postintervention and to see whether the type of control group used moderated the effects obtained. This meta-analysis is representative of clinical research in that relatively few studies had addressed this question (it is a small analysis) and sample sizes within each study were relatively small. These data are used as our main example, and the most benefit can be gained from reading the original meta-analysis in conjunction with this chapter. We will now look at each stage of the process of doing a meta-analysis.

Step 1: Do a Literature Search

The first step is to search the literature for studies that have addressed the core/central/same research question using electronic databases such as the ISI Web of Knowledge, PubMed, and PsycInfo. Although the obvious reason for doing this is to find articles, it is also helpful in identifying authors who might have unpublished data (see below). It is often useful to hand-search the reference sections of the articles that you have found to check for articles that you have missed, and to consult directly with noted experts in this literature to ensure that relevant papers have not been missed.

Although it is tempting to assume that meta-analysis is a wonderfully objective tool, it is not without a dash of bias. Possibly the main source of bias is the "file-drawer problem," or publication bias (Rosenthal, 1979). This bias stems from the reality that significant findings are more likely to be published than nonsignificant findings: significant findings are estimated to be eight times more likely to be submitted than nonsignificant ones (Greenwald, 1975), studies with positive findings are around seven times more likely to be published than studies with results supporting the null hypothesis (Coursol & Wagner, 1986), and 97 percent of articles in psychology journals report significant results (Sterling, 1959). Without rigorous attempts to counteract publication bias, meta-analytic reviews could overestimate population effects because effect sizes in unpublished studies will be smaller (McLeod & Weisz, 2004)—up to half the size (Shadish, 1992)—of published studies of comparable methodological quality. The best way to minimize the bias is to extend your search to relevant conference

proceedings and to contact experts in the field to see if they have or know of any unpublished data. This can be done by direct email to authors in the field, but also by posting a message to a topic-specific newsgroup or email listserv.

In our study, we gathered articles by searching PsycInfo, Web of Science, and Medline for English-language studies using keywords considered relevant. The reference lists of previous meta-analyses and retrieved articles were scanned for relevant studies. Finally, email addresses of published researchers were compiled from retrieved papers, and 52 researchers were emailed and invited to send any unpublished data fitting the inclusion criteria. This search strategy highlights the use of varied resources to ensure that all potentially relevant studies are included and to reduce bias due to the file-drawer problem.

Step 2: Decide on Inclusion Criteria

A second source of bias in a meta-analysis is the inclusion of poorly conducted research. As Field and Gillett (2010) put it:

> Although meta-analysis might seem to solve the problem of variance in study quality because these differences will "come out in the wash," even one red sock (bad study) amongst the white clothes (good studies) can ruin the laundry. (pp. 667–668)

Inclusion criteria depend on the research question being addressed and any specific methodological issues in the field, but the guiding principle is that you want to compare apples with apples, not apples with pears (Eysenck, 1978). In a meta-analysis of CBT, for example, you might decide on a working definition of what constitutes CBT, and maybe exclude studies that do not have proper control groups. It is important to use a precise, reliable, set of criteria that is applied consistently to each potential study so as not to introduce subjective bias into the analysis. In your write-up, you should be explicit about your inclusion criteria and report the number of studies that were excluded at each step of the selection process.

In our analysis, at a first pass we excluded studies based on the following criteria: (a) treatments were considered too distinct to be meaningfully compared to face-to-face therapies (e.g., bibliotherapy, telephone, or computer-administered treatment); (b) subsamples of the data were already included in the meta-analysis because they were published over several papers; and (c) information was insufficient to enable us to compute effect sizes. Within this pool of studies, we set the following inclusion criteria:

1. Studies that included only those participants who met criteria for a diagnosis of GAD outlined by the DSM since GAD was recognized as an independent disorder; that is, the DSM-III-R, DSM-IV, or DSM-IV-TR (prior to DSM-III-R, GAD was simply a poorly characterized residual diagnostic category). This was to avoid samples being heterogeneous.
2. Studies in which the majority of participants were aged 18 to 65 years. This was because there may be developmental issues that affect the efficacy of therapy in younger samples.
3. The Penn State Worry Questionnaire (PSWQ) was used to capture symptom change.
4. Treatments included were defined as any treatment that used cognitive techniques, either in combination with, or without, behavioral techniques.
5. To ensure that the highest possible quality of data was included, only studies that used a randomized controlled design were included.

Step 3: Calculate the Effect Sizes

WHAT ARE EFFECT SIZES AND HOW DO I CALCULATE THEM?

Your selected studies are likely to have used different outcome measures, and of course we cannot directly compare raw change on a children's self-report inventory to that being measured using a diagnostic tool such as the Anxiety Disorders Interview Schedule (ADIS). Therefore, we need to standardize the effects within each study so that they can be combined and compared. To do this we convert each effect in each study into one of many standard effect size measures. When quantifying group differences on a continuous measure (such as the PSWQ) people tend to favor Cohen's *d*; Pearson's *r* is used more when looking at associations between measures; and if recovery rates are the primary interest, then it is common to see odds ratios used as the effect size measure.

Once an effect size measure is chosen, you need to compute it for each effect that you want to compare for every paper you want to include in the meta-analysis. A given paper may contain several effect sizes depending on the sorts of questions you are trying to address with your meta-analysis. For example, in a meta-analysis on cognitive impairment in PTSD in which I was involved (Brewin et al., 2007), impairment was measured in a variety of

ways in individual studies, and so we had to compute several effect sizes within many of the studies. In this situation, we have to make a decision about how to treat the studies that have produced multiple effect sizes that address the same question. A common solution is to calculate the average effect size across all measures of the same outcome within a study (Rosenthal, 1991), so that every study contributes only one effect to the main analysis (as in Brewin et al., 2007).

Computing effect sizes is probably the hardest part of a meta-analysis because the data contained within published articles will vary in their detail and specificity. Some articles will report effect sizes, but many will not; articles might use different effect size metrics; you will feel as though some studies have a grudge against you and are trying to make it as hard for you as possible to extract an effect size. If no effect sizes are reported, then you need to try to use the reported data to calculate one. If using *d*, then you can use means and standard deviations, odds ratios are easily obtained from frequency data, and most effect size measures (including *r*) can be obtained from test statistics such as *t*, *z*, χ^2, and *F*, or probability values for effects (by converting first to *z*). A full description of the various ways in which effect sizes can be computed is beyond the present scope, but there are many freely available means to compute effect sizes from raw data and test statistics; some examples are Wilson (2001, 2004) and DeCoster (1998). To do the meta-analysis you need not just the effect size, but the corresponding value of its sampling variance (*v*) or standard error (*se*); Wilson (2001), for example, will give you an estimate of the effect size and the sampling variance.

If a paper does not include sufficient data to calculate an effect size, contact the authors for the raw data, or relevant statistics from which an effect size can be computed. (If you are on the receiving end of such an email please be sympathetic, as attempts to get data out of researchers can be like climbing up a jelly mountain.)

EFFECT SIZES FOR HANRAHAN AND COLLEAGUES' STUDY

When reporting a meta-analysis it is a good idea to tabulate the effect sizes with other helpful information (such as the sample size on which the effect size is based, *N*) and also to present a stem-and-leaf plot of the effect sizes. For the study conducted by Hanrahan and colleagues, we used *d* as the effect size measure and corrected for the known bias that *d* has in small samples using the adjustment described by Hedges (1981). In meta-analysis, a stem-and-leaf plot graphically organizes included effect sizes to visualize the shape and central tendency of the effect size distribution across studies included. Table 17.1 shows a stem-and-leaf plot of the resulting effect sizes, and this should be included in the write-up. This stem-and-leaf plot tells us the effect sizes to one decimal place, with the stem reflecting the value before the decimal point and the leaf showing the first decimal place; for example, we know the smallest effect size was $d = -0.2$, the largest was $d = 3.2$, and there were effect sizes of 1.2 and 1.4 (for example). Table 17.2 shows the studies included in the Hanrahan and colleagues' paper, with their corresponding effect sizes (expressed as *d*), the sample sizes on which these *d*s are based, and the standard errors associated with each effect size. Note that the *d*s match those reported in Table 2 of Hanrahan and colleagues (2013).

Table 17.1 Stem-and-Leaf Plot of All Effect Sizes (*d*s)

Stem	Leaf
−0	2, 1
0	1, 3, 3, 3, 3, 4, 7, 8, 8, 9
1	2, 4,
2	2, 2, 4, 6
3	2

Step 4: Do the Basic Meta-Analysis

INITIAL CONSIDERATIONS

Meta-analysis aims to estimate the effect in the population (and a confidence interval around it) by combining the effect sizes from different studies using a weighted mean of the effect sizes. The "weight" that is used is usually a value reflecting the sampling precision of the effect size, which is typically a function of sample size. As such, effect sizes with better precision are weighted more highly than effect sizes that are imprecise. There are different methods for estimating the population effects, and these methods have pros and cons. There are two related issues to consider: (1) which method to use and (2) how to conceptualize your data. There are other issues, too, but we will focus on these two because there are articles elsewhere that can be consulted as a next step (e.g., Field, 2001, 2003a, 2003b, 2005a, 2005b; Hall & Brannick, 2002; Hunter & Schmidt, 2004; Rosenthal & DiMatteo, 2001; Schulze, 2004).

Table 17.2 Effect Sizes (*d*) from Hanrahan and colleagues (2013)

Study ID	Study	ES ID	Control type	*N*	*d*	SE(d)
1	Van der Heiden et al. (under review)	1	Non-therapy	81	1.42	0.278
		2	CT	121	0.68	0.186
2	Newman et al. (under review)	3	CT	83	–0.17	0.218
3	Wells et al. (2010)	4	Non-CT	20	2.57	0.590
4	Dugas et al. (2010)	5	Non-therapy	43	0.82	0.313
		6	Non-CT	45	0.13	0.293
5	Westra et al. (2009)	7	CT	76	0.45	0.230
6	Leichsenring et al. (2009)	8	Non-CT	57	0.31	0.263
7	Roemer et al. (2008)	9	Non-therapy	31	2.44	0.468
8	Rezvan et al. (2008)	10	Non-therapy	24	3.22	0.609
		11	CT	24	–0.08	0.394
9	Zinbarg et al. (2007)	12	Non-therapy	18	2.25	0.587
10	Gosselin et al. (2006)	13	Non-CT	53	0.89	0.284
11	Dugas et al. (2003)	14	Non-therapy	52	1.23	0.299
12	Borkovec et al. (2002)	15	CT	46	0.27	0.291
13	Ladouceur et al. (2000)	16	Non-therapy	26	2.22	0.490
14	Ost & Breitholtz (2000)	17	Non-CT	33	0.31	0.323
15	Borkovec & Costello (1993)	18	Non-CT	37	0.83	0.336
		19	Non-CBT	37	0.26	0.323

CHOOSING A MODEL

It is tempting simply to tell you to use a random-effects model and end the discussion: however, in the interests of informed decision making I will explain why. Meta-analysis can be conceptualized in two ways: fixed- and random-effects models (Hedges, 1992; Hedges & Vevea, 1998; Hunter & Schmidt, 2000). We can assume that studies in a meta-analysis are sampled from a population in which the average effect size is fixed (Hunter & Schmidt, 2000). Consequently, sample effect sizes should be homogenous. This is the fixed-effects model. The alternative assumption is that the average effect size in the population varies randomly from study to study: population effect sizes can be thought of as being sampled from a "superpopulation" (Hedges, 1992). In this case, because effect sizes come from populations with varying average effect sizes, they should be heterogeneous. This is the random-effects model. Essentially, the researcher using a random-effects model assumes that the studies included represent a mere random sampling of the larger population of studies that *could* have been conducted on the topic, whereas the researcher using a fixed-effects model assumes that the studies included are the comprehensive set of representative studies. Therefore, the fixed-effects model can be thought to characterize the scope of existing research, and the random-effects model can be thought to afford inferences about a broader population than just the sample of studies analyzed. When effect size variability is explained by a moderator variable that is treated as "fixed," then the random-effects model becomes a mixed-effects model (see Overton, 1998).

Statistically speaking, fixed- and random-effects models differ in the sources of error. Fixed-effects models have error derived from sampling studies from a population of studies. Random-effects models have this error too, but in addition there is error created by sampling the populations from a superpopulation.

The two most widely used methods of meta-analysis are those by Hunter and Schmidt (2004),

which is a random-effects method, and the method by Hedges and colleagues (e.g., Hedges, 1992; Hedges & Olkin, 1985; Hedges & Vevea, 1998), who provide both fixed- and random-effects methods. However, multilevel models can also be used in the context of meta-analysis (see Hox, 2002, Chapter 8).

Your first decision is whether to conceptualize your model as fixed- or random-effects. You might consider the assumptions that can be realistically made about the populations from which your studies are sampled. There is compelling evidence that real-world data in the social sciences are likely to have variable population parameters (Field, 2003b; Hunter & Schmidt, 2000, 2004; National Research Council, 1992; Osburn & Callender, 1992). Field (2005a) found that the standard deviations of effect sizes for all meta-analytic studies (using *r*) published in *Psychological Bulletin* from 1997 to 2002 ranged from 0 to 0.3 and were most frequently in the region of 0.10 to 0.16; similarly, Barrick and Mount (1991) reported that the standard deviation of effect sizes (*r*s) in published datasets was around 0.16.

Second, consider the inferences that you wish to make (Hedges & Vevea, 1998): if you want to make inferences that extend only to the studies included in the meta-analysis (*conditional inferences*), then fixed-effect models are appropriate; however, for inferences that generalize beyond the studies in the meta-analysis (*unconditional inferences*), a random-effects model is appropriate.

Third, consider the consequences of making the "wrong" choice. The consequences of applying fixed-effects methods to random-effects data can be quite dramatic: (1) it inflates the significance tests of the estimate of the population effect from the normal 5 percent to 11 to 28 percent (Hunter & Schmidt, 2000) and 43 to 80 percent (Field, 2003b) and (2) published fixed-effects confidence intervals around mean effect sizes have been shown to be, on average, 52 percent narrower than their actual width—these nominal 95 percent fixed-effects confidence intervals were on average 56 percent confidence intervals (Schmidt, Oh, & Hayes, 2009). The consequences of applying random-effects methods to fixed-effects data are considerably less dramatic: in Hedges' method, for example, when sample effect sizes are homogenous, the additional between-study effect size variance becomes zero, yielding the same result as the fixed-effects method.

This leads me neatly back to my opening sentence of this section: unless you can find a good reason not to, use a random-effects method because (1) social science data normally have heterogeneous effect sizes; (2) psychologists generally want to make inferences that extend beyond the studies in the meta-analysis; and (3) if you apply a random-effects method to homogenous effect sizes, it does not affect the results (certainly not as dramatically as if you apply a fixed-effects model to heterogeneous effect sizes).

CHOOSING A METHOD

Let's assume that you trust me (I have an honest face) and opt for a random-effects model. You then need to decide whether to use Hunter and Schmidt, H-S (2004) or Hedges and colleagues' method (H-V). The technical differences between these methods have been summarized elsewhere (Field, 2005a) and will not be repeated here. In a series of Monte Carlo simulations comparing the performance of the Hunter and Schmidt and Hedges and Vevea (fixed- and random-effects) methods, Field (2001; but see Hafdahl & Williams, 2009) found that when comparing random-effects methods, the Hunter-Schmidt method yielded the most accurate estimates of population correlations across a variety of situations (a view echoed by Hall & Brannick, 2002, in a similar study). Based on a more extensive set of stimulations, Field (2005a) concluded that in general both H-V and H-S random-effects methods produce accurate estimates of the population effect size. Although there were subtle differences in the accuracy of population effect size estimates across the two methods, in practical terms the bias in both methods was negligible. In terms of 95 percent confidence intervals around the population estimate, Hedges' method was in general better at achieving these intervals (the intervals for Hunter and Schmidt's method tended to be too narrow, probably because they recommend using credibility intervals and not confidence intervals).

Hunter and Schmidt's method involves psychometric corrections for the attenuation of observed effect sizes that can be caused by measurement error (Hunter, Schmidt, & Le, 2006), and these psychometric corrections can be incorporated into the H-V method if correlations are used as the effect size, but these corrections were not explored in the studies mentioned above, which limits what they can tell us. Therefore, diligent researchers might consult the various tables in Field (2005a) to assess which method might be most accurate for the given parameters of the meta-analysis that they are about to conduct; however, the small differences between the methods will probably not make a substantive

impact on the conclusions that will be drawn from the analysis.

ENTERING THE DATA INTO *R*

Having computed the effect sizes, we need to enter these into *R*. In *R*, commands follow the basic structure of:

```
Object <- Instructions about how to create the object
```

Therefore, to create an object that is a variable, we give it a name on the left-hand side of the arrow, and on the right-hand side input the data that makes up the variable. To input data we use the *c()* function, which simply binds things together into a single object (in this case it binds the different values of *d* into a single object or variable). To enter the value of *d* from Table 17.2, we would execute:

```
d <- c(1.42, 0.68, –0.17, 2.57, 0.82, 0.13, 0.45,
0.31, 2.44, 3.22, –0.08, 2.25, 0.89, 1.23, 0.27, 2.22,
0.31, 0.83, 0.26)
```

Executing this command creates an object that we have named "d" (we could have named it "Thelma" if we wanted to, but "d" seems like a fairly descriptive name in the circumstances). If we want to view this variable, we simply execute its name:

```
> d
[1] 1.42 0.68 –0.17 2.57 0.82 0.13 0.45 0.31
2.44 3.22 –0.08 2.25 0.89 1.23 0.27 2.22 0.31
0.83 0.26
```

We can enter the standard errors from Table 17.2 in a similar way; this time we create an object that we have decided to call "sed":

```
sed <- c(0.278, 0.186, 0.218, 0.590, 0.313, 0.293,
0.230, 0.263, 0.468, 0.609, 0.394, 0.587, 0.284,
0.299, 0.291, 0.490, 0.343, 0.336, 0.323)
```

Next I'm going to create a variable that gives each effect size a label of the first author of the study from which the effect came and the year. We can do this by executing (note that to enter text strings instead of numbers, we place the text in quotes so that *R* knows the data are text strings):

```
study <- c("v.d. Heiden (2010)", "v.d. Heiden
(2010)", "Newman (2010)", "Wells (2010)",
"Dugas (2010)", "Dugas (2010)", "Westra (2009)",
"Leichsenring (2009)", "Roemer (2008)", "Rezvan
(2008)", "Rezvan (2008)", "Zinbarg (2007)",
"Gosselin (2006)", "Dugas (2003)", "Borkovec
(2002)", "Ladouceur (2000)", "Ost (2000)",
"Borkovec (1993)", "Borkovec (1993)")
```

We also have some information about the type of control group used. We'll come back to this later, but if we want to record this information, we can do so using a coding variable. We need to enter values that represent the different types of control group, and then to tell *R* what these values represent. Let's imagine that we want to code non-therapy controls as 0, CT as 1, and non-CT as 2. First we can enter these values into *R*:

```
controlType <- c(0, 1, 1, 2, 0, 2, 1, 2, 0, 0, 1, 0, 2, 0,
1, 0, 2, 2, 2)
```

Next, we need to tell *R* that this variable is a coding variable (a.k.a. a factor), using the *factor()* function. Within this function we name the variable that we want to convert (in this case *controlType*), we tell *R* what numerical values we have used to code levels of the factor by specifying *levels = 0:2* (0:2 means zero to 2 inclusive, so we are specifying levels of 0, 1, 2), we then tell it what labels to attach to those levels (in the order of the numerical codes) by including *labels = c("Non-Therapy", "CT", "Non-CT")*. Therefore, to turn *controlType* into a factor based on itself, we execute:

```
controlType <- factor(controlType, levels = 0:2,
labels = c("Non-Therapy", "CT", "Non-CT"))
```

We now have four variables containing data: *d* (the effect sizes), *sed* (their standard errors), *study* (a string variable that identifies the study from which the effect came), and *controlType* (a categorical variable that defines what control group was used for each effect size). We can combine these variables into a data frame by executing:

```
GAD.data <- data.frame(study, controlType, d, sed)
```

This creates an object called *GAD.data* (note that in *R* you cannot use spaces when you name objects) that is a data frame made up of the four variables that we have just created. To "see" this data frame, execute its name:

```
> GAD.data
study controlType d sed
1 v.d. Heiden (2010) Non-Therapy 1.42 0.278
2 v.d. Heiden (2010) CT 0.68 0.186
3 Newman (2010) CT -0.17 0.218
4 Wells (2010) Non-CT 2.57 0.590
5 Dugas (2010) Non-Therapy 0.82 0.313
6 Dugas (2010) Non-CT 0.13 0.293
7 Westra (2009) CT 0.45 0.230
8 Leichsenring (2009) Non-CT 0.31 0.263
9 Roemer (2008) Non-Therapy 2.44 0.468
```

10 Rezvan (2008) Non-Therapy 3.22 0.609
11 Rezvan (2008) CT -0.08 0.394
12 Zinbarg (2007) Non-Therapy 2.25 0.587
13 Gosselin (2006) Non-CT 0.89 0.284
14 Dugas (2003) Non-Therapy 1.23 0.299
15 Borkovec (2002) CT 0.27 0.291
16 Ladouceur (2000) Non-Therapy 2.22 0.490
17 Ost (2000) Non-CT 0.31 0.343
18 Borkovec (1993) Non-CT 0.83 0.336
19 Borkovec (1993) Non-CT 0.26 0.323

You can also prepare the data as a tab-delimited or comma-separated text file (using Excel, SPSS, Stata, or whatever other software you like) and read this file into *R* using the *read.delim()* or *read.csv()* functions. In both cases, you could use the *choose.file()* function to open a standard dialogue box that lets you choose the file by navigating your file system.[1] For example, to create a data frame called "GAD.data" from a tab-delimited file (.dat), you would execute:

```
GAD.data <- read.delim(file.choose(), header =
TRUE)
```

Similarly, to create a data frame from a comma-separated text file, you would execute:

```
GAD.data <- read.csv(file.choose(), header = TRUE)
```

In both cases the *header = TRUE* option is used if you have variable names in the first row of the data file; if you do not have variable names in the file, omit this option (the default value is false). If this data-entry section has confused you, then read Chapter 3 of Field and colleagues (2012) or your preferred introductory book on *R*.

DOING THE META-ANALYSIS

To do a basic meta-analysis you use the *rma()* function. This function has the following general format when using the standard error of effect sizes:

```
maModel <- rma(yi = variable containing effect sizes,
sei = variable containing standard error of effect sizes,
data = dataFrame, method = "DL")
```

maModel is whatever name you want to give your model (but remember you can't use spaces), *variable containing effect sizes* is replaced with the name of the variable containing your effect sizes, *variable containing standard error of effect sizes* is replaced with the name of the variable that contains the standard errors, and *dataFrame* is the name of the data frame containing these variables.

When using the variance of effect sizes we substitute the *sei* option with *vi*:

```
maModel <- rma(yi = variable containing effect sizes,
vi = variable containing variance of effect sizes, data =
dataFrame, method = "DL")
```

Therefore, for our GAD analysis, we can execute:

```
maGAD <- rma(yi = d, sei = sed, data = GAD.data,
method = "DL")
```

This creates an object called *maGAD* by using the *rma()* function. Within this function, we have told *R* that we want to use the object *GAD.data* as our data frame, and that within this data frame, the variable *d* contains the effect sizes (*yi = d*) and the variable *sed* contains the standard errors (*sei = sed*). Finally, we have set the method to be "DL," which will use the DerSimonian-Laird estimator (which is used in the H-V random-effects method). We can change how the model is estimated by changing this option, which can be set to the following:

- method = "FE": produces a fixed-effects meta-analysis
- method = "HS" = random effects using the Hunter-Schmidt estimator
- method = "HE" = random effects using the Hedges estimator
- method = "DL" = random effects using the DerSimonian-Laird estimator
- method = "SJ" = random effects using the Sidik-Jonkman estimator
- method = "ML" = random effects using the maximum-likelihood estimator
- method = "REML" = random effects using the restricted maximum-likelihood estimator (this is the default if you don't specify a method at all)
- method = "EB" = random effects using the empirical Bayes estimator.

To see the results of the analysis we need to use the *summary()* function and put the name of the model within it:

```
summary(maGAD)
```

The resulting output can be seen in Figure 17.2. This output is fairly self-explanatory[2]. For example, we can see that for Hedges and Vevea's method, the *Q* statistic, which measures heterogeneity in effect sizes, is highly significant, χ^2 (18) = 100.50, $p < .001$. The estimate of between-study variability, $\tau^2 = 0.44$ (most important, this is not zero), and the proportion of variability due to heterogeneity,

```
Random-Effects Model (k = 19; tau^2 estimator: DL)

  logLik  Deviance       AIC       BIC
-26.1284   52.2568   56.2568   58.1457

tau^2 (estimate of total amount of heterogeneity): 0.4358
tau (sqrt of the estimate of total heterogeneity): 0.6602
I^2 (% of total variability due to heterogeneity): 82.09%
H^2 (total variability / sampling variability):    5.58

Test for Heterogeneity:
Q(df = 18) = 100.5016, p-val < .0001

Model Results:

estimate       se     zval     pval    ci.lb    ci.ub
  0.9303   0.1723   5.4001   <.0001   0.5926   1.2679      ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 17.2 *R* output for a basic meta-analysis.

I^2, was 82.09 percent. In other words, there was a lot of variability in study effects. The population effect size and its 95 percent confidence interval are: 0.93 ($CI_{.95}$ = 0.59 (lower), 1.27 (upper)). We can also see that this population effect size is significant, $z = 5.40$, $p < .001$.

Based on the homogeneity estimates and tests, we could say that there was considerable variation in effect sizes overall. Also, based on the estimate of population effect size and its confidence interval, we could conclude that there was a strong effect of CT for GAD compared to controls.

Creating a forest plot of the studies and their effect sizes is very easy after having created the meta-analysis model. We simply place the name of the model within the *forest()* command and execute:

```
forest(maGAD)
```

However, I want to add the study labels to the plot, so let's execute:

```
forest(maGAD, slab = GAD.data$study)
```

By adding *slab = GAD.data$study* to the command we introduce study labels (that's what *slab* stands for) and the labels we use are in the variable called study within the *GAD.data* data frame (that's what *GAD.data$study* means). The resulting figure is in Figure 17.3. It shows each study with a square indicating the effect size from that study (the size of the square is proportional to the weight used in the meta-analysis, so we can see that the first three studies were weighted fairly heavily). The branches of each effect size represent the confidence interval of the effect size. Also note that because we added the *slab* option, our effects have been annotated using the names in the variable called *study* in our data frame. Looking at this plot, we can see that there are five studies that produced fairly substantially bigger effects than the rest, and two studies with effect sizes below zero (the dotted line), which therefore showed that CBT was worse than controls. The diamond at the bottom shows the population effect size based on these individual studies (it is the value of the population effect size from our analysis). The forest plot is a very useful way to summarize the studies in the meta-analysis.

Step 5: Do Some More Advanced Analysis

ESTIMATING PUBLICATION BIAS

Various techniques have been developed to estimate the effect of publication bias and to correct for it. The earliest and most commonly reported estimate of publication bias is Rosenthal's (1979) fail-safe *N*. This was an elegant and easily understood method for estimating the number of unpublished studies that would need to exist to turn a significant population effect size estimate into a nonsignificant one. However, because significance testing the estimate of the population effect size is not really the reason for doing a meta-analysis, the fail-safe *N* is fairly limited.

The funnel plot (Light & Pillerner, 1984) is a simple and effective graphical technique for exploring potential publication bias. A funnel plot displays effect sizes plotted against the sample size, standard error, conditional variance, or some other measure of the precision of the estimate. An unbiased sample would ideally show a cloud of data points that is symmetrical around the population effect size and has the shape of a funnel. This funnel shape reflects the greater variability in effect sizes from studies with small sample sizes/less precision, and the estimates drawn from larger/more precise studies converging

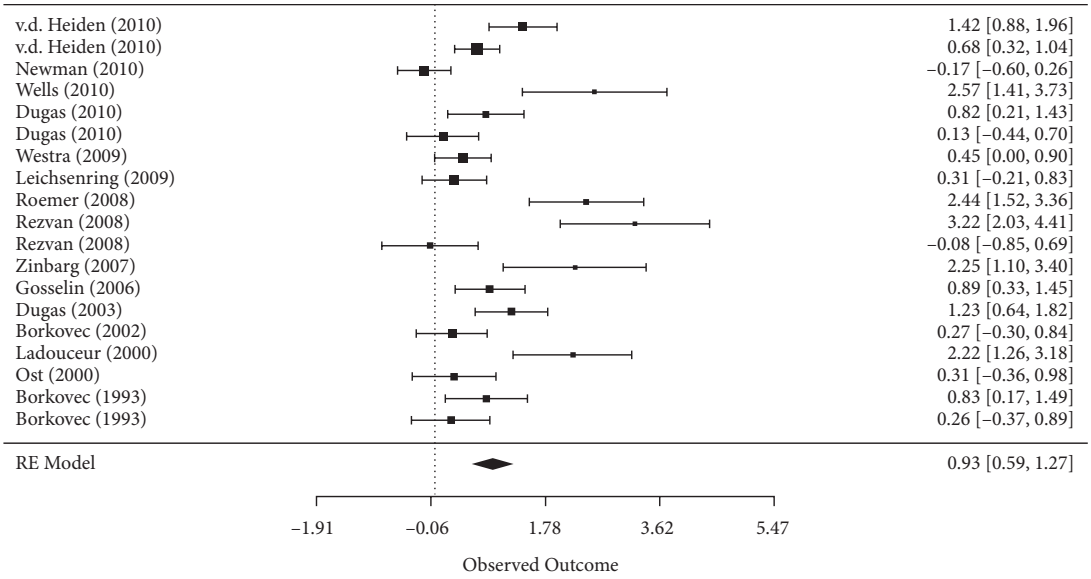


Figure 17.3 Forest plot of the GAD data.

around the population effect size. A sample with publication bias will lack symmetry because studies based on small samples that showed small effects will be less likely to be published than studies based on the same-sized samples that showed larger effects (Macaskill, Walter, & Irwig, 2001).

Funnel plots should be used as a first step before further analysis because factors other than publication bias can cause asymmetry. Some examples are data irregularities including fraud and poor study design (Egger, Smith, Schneider, & Minder, 1997), true heterogeneity of effect sizes (in intervention studies this can happen because the intervention is more intensely delivered in smaller, more personalized studies), and English-language bias (studies with smaller effects are often found in non–English-language journals and get overlooked in the literature search).

To get a funnel plot for a meta-analysis model created in *R*, we simply place that model into the *funnel()* function and execute:

```
funnel(maGAD)
```

Figure 17.4 shows the resulting funnel plot, which is clearly not symmetrical. The studies with large standard errors (bottom right) consistently produce the largest effect sizes, and the studies are not evenly distributed around the mean effect size (or within the unshaded triangle). This graph shows clear publication bias.

Funnel plots offer no means to correct for any bias detected. Trim and fill (Duval & Tweedie, 2000) is a method in which a biased funnel plot is truncated ("trimmed") and the number (*k*) of missing studies from the truncated part is estimated. Next, *k* artificial studies are added ("filled") to the negative side of the funnel plot (and therefore have small effect sizes) so that in effect the study now contains *k* new "studies" with effect sizes as small in magnitude as the *k* largest effect sizes that were trimmed. The new "filled" effects are presumed to represent the magnitude of effects identified in hypothetical unpublished studies. A new estimate of the population effect size is then calculated including these artificially small effect sizes. Vevea and Woods (2005) point out that this method can lead to overcorrection because it relies on the strict assumption that all of the "missing" studies are those with the smallest effect sizes. Vevea and Woods propose a more sophisticated correction method based on weight function models of publication bias. These methods use weights to model the process through which the likelihood of a study being published varies (usually based on a criterion such as the significance of a study). Their method can be applied to even small meta-analyses and is relatively flexible in allowing meta-analysts to specify the likely conditions of publication bias in their particular research scenario. (The downside of this flexibility is that it can be hard to know what the precise conditions are.) They specify four typical weight functions: "moderate one-tailed selection," "severe one-tailed selection," "moderate two-tailed selection," and "severe two-tailed selection"; however, they recommend adapting the weight functions based on what the funnel plot reveals (see Vevea &

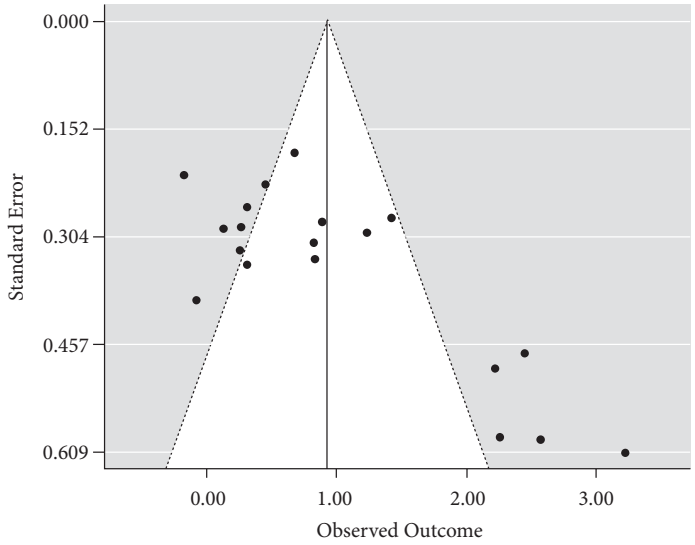


Figure 17.4 Funnel plot of the GAD data.

Woods, 2005). These corrections can be applied in *R* (see Field & Gillett, 2010, for a tutorial) but do not form part of the *metafor* package and are a little too technical for this introductory chapter.

MODERATOR ANALYSIS

When there is variability in effect sizes, it can be useful to try to explain this variability using theoretically driven predictors of effect sizes. For example, in our dataset there were three different types of control group used: non-therapy (waitlist), non-CT therapy, and CT therapy. We might reasonably expect effects to be stronger if a waitlist control was used in the study compared to a CT control because the waitlist control gets no treatment at all, whereas CT controls get some treatment. We can test for this using a mixed model (i.e., a random-effects model in which we add a fixed effect).

Moderator models assume a general linear model in which each effect size can be predicted from the moderator effect (represented by β_1):

$$ES = \beta_0 + C\beta_1 + e$$

The within-study error variance is represented by e_i. To calculate the moderator effect, β_1, a generalized least squares (GLS) estimate is calculated. It is not necessary to know the mathematics behind the process (if you are interested, then read Field, 2003b; Overton, 1998); the main thing to understand is that we're just doing a regression in which effect sizes are predicted. Like any form of regression we can, therefore, predict effect sizes from either continuous variables (such as study quality) or categorical ones (which will be dummy coded using contrast weights).

The package *metafor* allows both continuous and categorical predictors (moderators) to be entered into the regression model that a researcher wishes to test. Moderator variables are added by including the *mods* option to the basic meta-analysis command. You can enter a single moderator by specifying *mods* = *variableName* (in which *variableName* is the name of the moderator variable that you want to enter into the model) or enter several moderator variables by including *mods* = *matrixName* (in which *matrixName* is the name of a matrix that contains values of moderator variables in each of its columns). Continuous variables are treated as they are; for categorical variables, you should either dummy code them manually or use the *factor()* function, as we did earlier, in which case R does the dummy coding for you.

Therefore, in our example, we can add the variable *controlType* as a moderator by rerunning the model including *mods* = *controlType* into the command. This variable is categorical, but because we converted it to a factor earlier on, R will treat it as a dummy-coded categorical variable. The rest of the command is identical to before:

```
modGAD <- rma(yi = d, sei = sed, data = GAD.data,
mods = controlType, method = "DL")
summary(modGAD)
```

The resulting output is shown in Figure 17.5. This output is fairly self-explanatory; for example,

```
Mixed-Effects Model (k = 19; tau^2 estimator: DL)

  logLik  Deviance       AIC       BIC
-22.4142  44.8285   50.8285   53.6618

tau^2 (estimate of residual amount of heterogeneity): 0.3313
tau (sqrt of the estimate of residual heterogeneity): 0.5756

Test for Residual Heterogeneity:
QE(df = 17) = 76.3534, p-val < .0001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 8.9309, p-val = 0.0028

Model Results:

         estimate      se     zval    pval    ci.lb    ci.ub
intrcpt  2.0417      0.4098  4.9817  <.0001   1.2385   2.8450 ***
mods    -0.5540      0.1854 -2.9885  0.0028  -0.9173  -0.1907 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 17.5 Output from *R* for moderation analysis.

we can see that for Hedges and Vevea's method, the estimate of between-study variability, $\tau^2 = 0.33$, is less than it was before (it was 0.44), which means that our moderator variable has explained some variance. However, there is still a significant amount left to explain, $\chi^2\,(17) = 76.35$, $p < .001$.

The *Q* statistic shows that the amount of variance explained by *controlType* is highly significant, $\chi^2\,(1) = 8.93$, $p = .0028$. In other words, it is a significant predictor of effect sizes. The beta parameter for the moderator and its 95 percent confidence interval are: −0.55, $CI_{.95} = -0.92$ (lower), −0.19 (upper). We can also see that this parameter is significant, $z = -3.11$, $p = .0028$ (note that the *p* value is identical to the *Q* statistic because they're testing the same thing). In a nutshell, then, the type of control group had a significant impact on the effect that CT had on GAD (measured by the PSWQ). We could break this effect apart by running the main meta-analysis on the three control groups separately.

Step 6: Write It Up

There are several detailed guidelines on how to write up a meta-analysis. For clinical trials, the QUORUM and PRISMA guidelines are particularly useful (Moher et al., 1999; Moher, Liberati, Tetzlaff, Altman, & Grp, 2009), and more generally the American Psychological Association (APA) has published its own Meta-Analysis Reporting Standards (MARS; H. Cooper, Maxwell, Stone, Sher, & Board, 2008). In addition, there are individual articles that offer advice (e.g., Field & Gillett, 2010; Rosenthal, 1995). There is a lot of overlap in these guidelines, and Table 17.3 assimilates them in an attempt to give a thorough overview of the structure and content of a meta-analysis article. This table should need no elaboration, but it is worth highlighting some of the key messages:

1. *Introduction*: Be clear about the rationale for the meta-analysis: What are the theoretical, practical, or policy drivers of the research? Why is a meta-analysis necessary? What hypotheses do you have?
2. *Methods*: Be clear about your search and inclusion criteria. How did you reach the final sample of studies? The PRISMA guidelines suggest including a flowchart of the selection process: Figure 17.6 shows the type of flowchart suggested by PRISMA, which outlines the number of studies retained and eliminated at each phase of the selection process. Also, state your computational and analytic methods in sufficient detail: Which effect size measure are you using (and did you have any issues in computing these)? Which meta-analytic technique did you apply to the data and why? Did you do a subgroup or moderator analysis?
3. *Results*: Include a graphical summary of the effect sizes included in the study. A forest plot is a very effective way to show the reader the raw data. When there are too many studies for a forest plot, consider a stem-and-leaf plot. A summary table of studies and any important study characteristics/moderator variables is helpful. If you have carried out a moderator analysis, then you might also provide stem-and-leaf plots or forest plots for subgroups of the analysis. Always report statistics relating to the variability of effect sizes (these should include the actual estimate of variability as well as statistical tests of variability), and obviously the estimate of the population effect size and

Table 17.3 Summary and Assimilation of Various Guidelines for Reporting Meta-analysis

Heading	Subheading	Advice
Title		Identify the article as a meta-analysis.
Abstract	Objectives	Describe the research question explicitly.
	Data sources	Describe the databases searched/data-gathering strategy.
	Review methods	Describe the selection criteria, methods for validity assessment, effect size computation, study characteristics (e.g., type of participants in the primary studies), and details of analysis (see below for detail).
	Results	Report the characteristics of the studies included and excluded; the main results (e.g., population estimates and confidence intervals), subgroup analyses, and publication bias analysis.
	Conclusion	Emphasize the key take-home message for theory, policy, and/or practice within the context of any limitations.
Introduction		Define the research problem and the rationale and scope of the review. This could include an historical background; the main theoretical, policy, or practical drivers of the research question; the need for a meta-analysis; the rationale for any moderator variables; the population to which the problem applies; the types of studies and measures typically used and their strengths and weaknesses; a statement of any *a priori* hypothesis.
Methods	Search Strategy	Describe the information sources explored (databases, registers, personal files, expert informants, agencies, hand-searching), the keywords and search terms used, and any restrictions on the search (years considered, language of publication, etc.).
	Selection	State the inclusion and exclusion criteria specifically. These could include eligible design features (random assignment, minimum sample size?), minimum standard for the measures used, etc.
	Study coding	State how validity was assessed (i.e., multiple researchers assessing manuscripts, interrater reliability on exclusion/inclusion of studies).
	Data abstraction	Describe how effect sizes were computed, what information was used from studies, what was done when the necessary data were not available, etc.
	Study characteristics	Include a summary of key measures/outcomes, participant characteristics across studies included, how homogeneity of methods/interventions across studies was assessed.
	Details of analysis	Be clear about the effect size metric used; the method of combining results (e.g., random or fixed effects, software used, etc.); how missing data were handled; how studies contributing multiple effect sizes were handled; how effect size heterogeneity was assessed. Include a rationale for any subgroup analyses. How was publication bias assessed?
Results	Trial flow	Consider including flow diagram of the decision-making process. Alternatively, a statement about the number of citations examined, how many were excluded, and the reasons why.
	Study characteristics	Present descriptive data for each study (e.g., age, sample size, experiment/intervention details, important study features or moderators such as measures used, dose of drug/intervention, duration on study/intervention, follow-up period, type of control group, etc.).
	Quantitative data synthesis	Report agreement statistics for the assessment of the validity of the study selection process; present summary statistics overall and for subgroups: population effect size estimate, effect size variability, confidence intervals, parameter estimates for moderators, etc. Include funnel or forest plots. Report any sensitivity/publication bias results.
Discussion		Summarize the key findings; discuss clinical/theoretical inferences based on internal and external validity; discuss the likely impact of potential biases in the review process (e.g., publication bias); make a clear statement about the future research agenda.

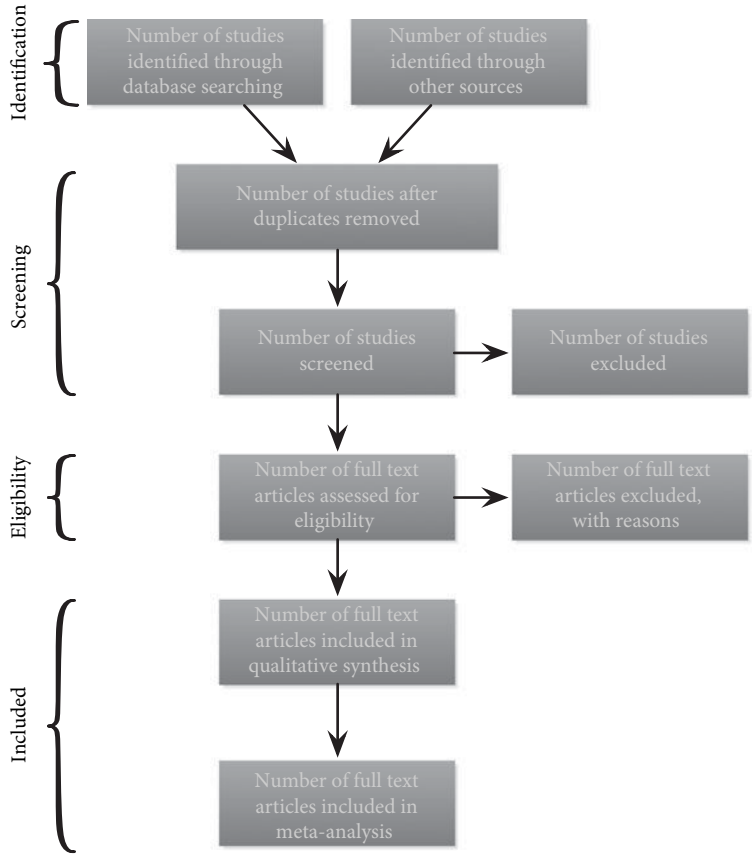


Figure 17.6 The PRISMA-recommended flowchart.

its associated confidence interval (or credibility interval). Report information on publication bias (e.g., a forest plot) and preferably a sensitivity analysis (e.g., Vevea and Woods' method).

4. *Discussion*: Pull out the key theoretical, policy, or practical messages that emerge from the analysis. Discuss any limitations or potential sources of bias within the analysis. Finally, it is helpful to make a clear statement about how the results inform the future research agenda.

Summary

This chapter offered a preliminary but comprehensive overview of the main issues when conducting a meta-analysis. We also used some real data to show how the *metafor* package in *R* can be used to conduct the analysis. The analysis begins by collecting articles about the research question you are trying to address through a variety of methods: emailing people in the field for unpublished studies, electronic searches, searches of conference abstracts, and so forth. Next, inclusion criteria should be devised that reflect the concerns pertinent to the particular research question (which might include the type of control group used, diagnostic measures, quality of outcome measure, type of treatment used, or other factors that ensure a minimum level of research quality). Statistical details are then extracted from the papers from which effect sizes can be calculated; the same effect size metric should be used for all studies, and you need to compute the variance or standard error for each effect size too. Choose the type of analysis appropriate for your particular situation (fixed- vs. random-effects, Hedges' method or Hunter and Schmidt's, etc.), and then apply this method to the data. Describe the effect of publication bias descriptively (e.g., funnel plots), and consider investigating how to re-estimate the population effect under various publication-bias models using Vevea and Woods' (2005) model. Finally, when reporting the results, make sure that the reader has clear information about the distribution of effect sizes (e.g., a stem-and-leaf plot), the effect size variability, the estimate of the population effect and its 95 percent confidence interval, the extent of publication bias, and whether any moderator variables were explored.

Useful Web Links

- Comprehensive Meta-Analysis: http://www.meta-analysis.com/
- MetaEasy: http://www.statanalysis.co.uk/meta-analysis.html
- *metafor* package for *R*: http://www.metafor-project.org/
- Mix: http://www.meta-analysis-made-easy.com/
- PRISMA (guidelines and checklists for reporting meta-analysis): http://www.prisma-statement.org/
- *R*: http://www.r-project.org/
- Review Manager: http://ims.cochrane.org/revman
- SPSS (materials accompanying Field & Gillett, 2010): http://www.discoveringstatistics.com/meta_analysis/how_to_do_a_meta_analysis.html

Notes

1. I generally find it easier to export from SPSS to a tab-delimited file because this format can also be read by software packages other than R. However, you can read SPSS data files (.sav) into R directly using the *read.spss()* function, but you need to first install and load a package called *foreign*.

2. There are slight differences in the decimal places between the results reported here and those on page 125 of Hanrahan and colleagues' papers because we did not round effect sizes and their standard errors to 2 and 3 decimal places respectively before conducting the analysis.

References

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133*(1), 1–24. doi: 10.1037/0033-2909.133.1.1

Barbato, A., & D'Avanzo, B. (2008). Efficacy of couple therapy as a treatment for depression: A meta-analysis. *Psychiatric Quarterly, 79*(2), 121–132. doi: 10.1007/s11126-008-9068-0

Barrick, M. R., & Mount, M. K. (1991). The big 5 personality dimensions and job-performance—a meta-analysis. *Personnel Psychology, 44*(1), 1–26.

Bax, L., Yu, L. M., Ikeda, N., Tsuruta, H., & Moons, K. G. M. (2006). Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology, 6*(50). http://www.biomedcentral.com/1471-2288/6/50

Bloch, M. H., Landeros-Weisenberger, A., Rosario, M. C., Pittenger, C., & Leckman, J. F. (2008). Meta-analysis of the symptom structure of obsessive-compulsive disorder. *American Journal of Psychiatry, 165*(12), 1532–1542. doi: 10.1176/appi.ajp.2008.08020320

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis* (Version 2). Englewood, NJ: Biostat. Retrieved from http://www.meta-analysis.com/

Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. *American Journal of Psychiatry, 162*(2), 214–227. doi: 10.1176/appi.ajp.162.2.214

Brewin, C. R., Kleiner, J. S., Vasterling, J. J., & Field, A. P. (2007). Memory for emotionally neutral information in posttraumatic stress disorder: A meta-analytic investigation. *Journal of Abnormal Psychology, 116*(3), 448–463. doi: Doi 10.1037/0021-843x.116.3.448

Burt, S. A. (2009). Rethinking environmental contributions to child and adolescent psychopathology: a meta-analysis of shared environmental influences. *Psychological Bulletin, 135*(4), 608–637. doi: 10.1037/a0015702

Cartwright-Hatton, S., Roberts, C., Chitsabesan, P., Fothergill, C., & Harrington, R. (2004). Systematic review of the efficacy of cognitive behaviour therapies for childhood and adolescent anxiety disorders. *British Journal of Clinical Psychology, 43*, 421–436.

Chan, R. C. K., Xu, T., Heinrichs, R. W., Yu, Y., & Wang, Y. (2010). Neurological soft signs in schizophrenia: a meta-analysis. *Schizophrenia Bulletin, 36*(6), 1089–1104. doi: 10.1093/schbul/sbp011

Cooper, H., Maxwell, S., Stone, A., Sher, K. J., & Board, A. P. C. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839–851.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: a step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology, 17*, 136–137.

Covin, R., Ouimet, A. J., Seeds, P. M., & Dozois, D. J. A. (2008). A meta-analysis of CBT for pathological worry among clients with GAD. *Journal of Anxiety Disorders, 22*(1), 108–116. doi: 10.1016/j.janxdis.2007.01.002

Crawley, M. J. (2007). *The R book*. Chichester: Wiley-Blackwell.

Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review, 30*(6), 768–778. doi: 10.1016/j.cpr.2010.06.001

DeCoster, J. (1998). *Microsoft Excel spreadsheets: Meta-analysis*. Retrieved October 1, 2006, from http://www.stat-help.com/spreadsheets.html

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634.

Eysenck, H. J. (1978). Exercise in mega-silliness. *American Psychologist, 33*(5), 517–517.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*(2), 161–180.

Field, A. P. (2003a). Can meta-analysis be trusted? *Psychologist, 16*(12), 642–645.

Field, A. P. (2003b). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2*, 77–96.

Field, A. P. (2005a). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10*(4), 444–467.

Field, A. P. (2005b). Meta-analysis. In J. Miles & P. Gilbert (Eds.), *A handbook of research methods in clinical and health psychology* (pp. 295–308). Oxford: Oxford University Press.

Field, A. P. (2009). Meta-analysis. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 404–422). London: Sage.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical & Statistical Psychology, 63*, 665–694.

Field, A. P., Miles, J. N. V., & Field, Z. C. (2012). *Discovering statistics using R: And sex and drugs and rock 'n' roll*. London: Sage.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). London: Oliver & Boyd.

Furr, J. M., Corner, J. S., Edmunds, J. M., & Kendall, P. C. (2010). Disasters and youth: a meta-analytic examination of posttraumatic stress. *Journal of Consulting and Clinical Psychology, 78*(6), 765–780. doi: 10.1037/A0021482

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Greenwald, A. G. (1975). Consequences of prejudice against null hypothesis. *Psychological Bulletin, 82*(1), 1–19.

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods, 14*(1), 24–42. doi: 10.1037/a0014697

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87*(2), 377–389.

Hanrahan, F., Field, A. P., Jones, F., & Davey, G. C. L. (2013). A meta-analysis of cognitive-behavior therapy for worry in generalized anxiety disorder. *Clinical Psychology Review*, *33*, 120–132..

Hedges, L. (1981). Distribution Theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17*(4), 279–296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Hendriks, G. J., Voshaar, R. C. O., Keijsers, G. P. J., Hoogduin, C. A. L., & van Balkom, A. J. L. M. (2008). Cognitive-behavioural therapy for late-life anxiety disorders: a systematic review and meta-analysis. *Acta Psychiatrica Scandinavica, 117*(6), 403–411. doi: 10.1111/j.1600-0447.2008.01190.x

Hox, J. J. (2002). *Multilevel analysis, techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*(4), 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*(3), 594–612. doi: 10.1037/0021-9010.91.3.594

Kashdan, T. B. (2007). Social anxiety spectrum and diminished positive experiences: Theoretical synthesis and meta-analysis. *Clinical Psychology Review, 27*(3), 348–365. doi: 10.1016/j.cpr.2006.12.003

Kleinstaeuber, M., Witthoeft, M., & Hiller, W. (2011). Efficacy of short-term psychotherapy for multiple medically unexplained physical symptoms: A meta-analysis. *Clinical Psychology Review, 31*(1), 146–160. doi: 10.1016/j.cpr.2010.09.001

Kontopantelis, E., & Reeves, D. (2009). MetaEasy: A meta-analysis add-in for Microsoft Excel. *Journal of Statistical Software, 30*(7). http://www.jstatsoft.org/v30/i07/paper

Lavesque, R. (2001). *Syntax: meta-analysis*. Retrieved October 1, 2006, from http://www.spsstools.net/

Light, R. J., & Pillerner, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20*(4), 641–654.

Malouff, J. A., Thorsteinsson, E. B., Rooke, S. E., Bhullar, N., & Schutte, N. S. (2008). Efficacy of cognitive behavioral therapy for chronic fatigue syndrome: A meta-analysis. *Clinical Psychology Review, 28*(5), 736–745. doi: 10.1016/j.cpr.2007.10.004

McLeod, B. D., & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology, 72*(2), 235–251.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F., et al. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet, 354*(9193), 1896–1900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Grp, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *Journal of Clinical Epidemiology, 62*(10), 1006–1012. doi: 10.1016/J.Jclinepi.2009.06.005

Mychailyszyn, M.P., Brodman, D., Read, K.L., & Kendall, P.C. (2012). Cognitive-behavioral school-based interventions for anxious and depressed youth: A meta-analysis of outcomes. *Clinical Psychology: Science and Practice, 19*(2), 129–153.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, D.C.: National Academy Press.

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology, 77*(2), 115–122.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*(3), 354–379.

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 39*(3), 250–261. doi: 10.1016/j.jbtep.2007.07.007

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining tests of significance. *Biometrika, 30*, 134–148.

Quick, J. M. (2010). *The statistical analysis with R beginners guide*. Birmingham: Packt.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Roberts, N. P., Kitchiner, N. J., Kenardy, J., Bisson, J. I., & Psych, F. R. C. (2009). Systematic review and meta-analysis of multiple-session early interventions following traumatic

Eventse *American Journal of Psychiatry, 166*(3), 293–301. doi: 10.1176/appi.ajp.2008.08040590

Rosa-Alcazar, A. I., Sanchez-Meca, J., Gomez-Conesa, A., & Marin-Martinez, F. (2008). Psychological treatment of obsessive-compulsive disorder: A meta-analysis. *Clinical Psychology Review, 28*(8), 1310–1325. doi: 10.1016/j.cpr.2008.07.001

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*(1), 185–193.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*(2), 183–192.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: the first 345 studies. *Behavioral and Brain Sciences, 1*(3), 377–386.

Ruocco, A. C. (2005). The neuropsychology of borderline personality disorder: A meta-analysis and review. *Psychiatry Research, 137*(3), 191–202. doi: 10.1016/j.psychres.2005.07.004

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical & Statistical Psychology, 62*, 97–128. doi: 10.1348/000711007x255327

Schulze, R. (2004). *Meta-analysis: a comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

Schwarzer, G. (2005). *Meta*. Retrieved October 1, 2006, from http://www.stats.bris.ac.uk/R/

Shadish, W. R. (1992). Do family and marital psychotherapies change what people do? A meta-analysis of behavioural outcomes. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 129–208). New York: Sage.

Singh, S. P., Singh, V., Kar, N., & Chan, K. (2010). Efficacy of antidepressants in treating the negative symptoms of chronic schizophrenia: meta-analysis. *British Journal of Psychiatry, 197*(3), 174–179. doi: 10.1192/bjp.bp.109.067710

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752–760.

Spreckley, M., & Boyd, R. (2009). Efficacy of applied behavioral intervention in preschool children with autism for improving cognitive, language, and adaptive behavior: a systematic review and meta-analysis. *Journal of Pediatrics, 154*(3), 338–344. doi: 10.1016/j.jpeds.2008.09.012

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*(285), 30–34.

Stewart, R. E., & Chambless, D. L. (2009). Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology, 77*(4), 595–606. doi: 10.1037/a0016032

Stouffer, S. A. (1949). *The American soldier: Vol. 1. Adjustment during Army life*. Princeton, NJ: Princeton University Press.

The Cochrane Collaboration. (2008). *Review Manager (RevMan) for Windows: Version 5.0*. Copenhagen: The Nordic Cochrane Centre. Retrieved from http://www.cc-ims.net/revman/

Verzani, J. (2004). *Using R for introductory statistics*. Boca Raton, FL: Chapman & Hall.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428–443.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Villeneuve, K., Potvin, S., Lesage, A., & Nicole, L. (2010). Meta-analysis of rates of drop-out from psychosocial treatment among persons with schizophrenia spectrum disorder. *Schizophrenia Research, 121*(1-3), 266–270. doi: 10.1016/j.schres.2010.04.003

Wilson, D. B. (2001). *Practical meta-analysis effect size calculator*. Retrieved August 3, 2010, from http://gunston.gmu.edu/cebcp/EffectSizeCalculator/index.html

Wilson, D. B. (2004). *A spreadsheet for calculating standardized mean difference type effect sizes*. Retrieved October 1, 2006, from http://mason.gmu.edu/~dwilsonb/ma.html

Zuur, A. F., Ieno, E. N., & Meesters, E. H. W. G. (2009). *A beginner's guide to R*. New York: Springer.

CHAPTER

18 Item Response Theory

Lynne Steinberg *and* David Thissen

Abstract

Item response theory (IRT) comprises a collection of mathematical and statistical models and methods that are used in the assembly and scoring of psychological questionnaires and scales. IRT is used to investigate the statistical properties of categorical responses to questions, or other observations that may be indicators on a test or scale. IRT models are used for item analysis and scoring for items with dichotomous or polytomous responses. Statistical analysis using IRT summarizes the degree of endorsement or severity of each response, the strength of the relation between the item response and the underlying construct, and the degree to which a collection of questions or other indicators measure one coherent construct, or more than one. IRT is also used to investigate the degree to which item responses exhibit differential item functioning, or a lack of desired invariance, over groups. Finally, IRT is used to compute scale scores that are comparable across alternate forms of a questionnaire, and that may have better statistical properties than more traditional summed scores. This chapter illustrates these ideas with empirical examples.

Key Words: Psychological measurement, test theory, psychometrics, item response theory, IRT, differential item functioning, DIF, item factor analysis

Introduction

In the context of the development of the forthcoming edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5), there has been discussion of an empirical model of psychopathology that incorporates dimensional personality traits in the classification of personality disorders (see, for example, Krueger & Eaton, 2010). *Dimensional* refers to a continuum that ranges from low to high and is intended to be in contrast to the categorical approach taken in the DSM-4 (American Psychiatric Association, 2000). Such a dimensional approach requires explicit measurement models for psychological constructs. The development of measurement instruments usually involves *items* with categorical responses; the items are conceptualized as indicators of level on the psychological construct. Such constructs may be broadly or narrowly defined. Item response theory (IRT) provides ways to assess the extent to which items measure individual differences on some specified construct. This chapter describes some aspects of IRT and provides illustrations of its use to measure dimensional personality traits.

IRT comprises a collection of mathematical and statistical models and methods that are used for item analysis, to construct tests and questionnaires, and to compute scores that have different properties than simple summed scores. Most often, the items that are analyzed using IRT are questions with scored responses. However, any set of indicators of some construct may be the "items" if each indicator yields some small number of categorical responses. Examples of indicators that are not verbal questions include scored behavioral observations and rescored combinations of responses that are sometimes called *testlets* (small tests within a test).

The application of IRT to data involves statistical estimation of the parameters of suitable item response models, evaluation of the fit of the model to the data to the extent that is feasible, test or questionnaire assembly, and/or the development of scoring algorithms using the item models with their estimated parameters.

For item analysis, estimates of the parameters of IRT models provide information about the discriminating power of an item separately from its *difficulty* or *severity*, which, depending on context, are labels for parameters associated with the proportion endorsing each response. In that respect, IRT parameter estimates may be more useful than the statistics most often associated with the traditional test theory; those often confound discrimination with difficulty or severity. For test construction or questionnaire assembly, IRT provides information that can be used to achieve any of a number of distinct goals—overall reliability may be maximized, or precision of measurement may be optimized near score levels associated with one or more decisions, or equal precision across a wide range of individual differences may be sought. Scores may be made comparable across multiple forms of a test or questionnaire that include different items. An extreme example of multiple forms of a test or questionnaire that include different items arises in computerized adaptive testing, in which items are selected adaptively, depending on previous responses, to optimize measurement for each respondent.

Contemporary IRT is the culmination of over 80 years of intellectual history. As summarized by Thissen and Steinberg (2009, p. 154), the essential ideas of all IRT models are that:

- Items have parameters that place them on the same scale as the variable being measured (Thurstone, 1925).
- The variable being measured is *latent* (or unobserved) (Lazarsfeld, 1950; Lord, 1952).
- The unobserved variable accounts for the observed interrelationships among the item responses (Lazarsfeld, 1950).

The intellectual history of IRT models has been summarized by Bock (1997a), Thissen and Orlando (2001), and Thissen and Steinberg (2009); the interested reader is referred to those sources for theoretical background. Reise and Waller (2009) have provided an overview of the conceptual issues involved in the use of IRT in clinical measurement. In this chapter we illustrate the contemporary state of IRT with examples.

Models for Items with Dichotomous Responses

The Two-Parameter Logistic Model

Much of the early development of IRT involved the normal ogive model for dichotomous responses, based on theoretical ideas described by Lord and Novick (1968, pp. 370–371). However, Birnbaum (1968) suggested the use of the logistic function in place of the normal integral, because the logistic simplifies computations involved in maximum likelihood (ML) estimation of the parameters. For the past 40 years, most software for IRT analysis has used logistic models, and so have most applications. All of the illustrations in this chapter use logistic models.

A prototypical IRT model is the two-parameter logistic (2PL) model for dichotomous responses. The equation for a logistic *trace line* (Lazarsfeld, 1950, p. 364) that gives the probability of endorsement of item i is

$$T_i(u_i = 1) = \frac{1}{1 + \exp[-(a_i\theta + c_i)]}, \qquad (1)$$

in which u_i is the item response (1 for endorsement, 0 otherwise);

θ is the latent variable being measured; and

a_i and c_i are two item parameters (hence, the *two*-parameter logistic).

The a parameter is usually referred to as the *discrimination* or *slope* parameter. The slope parameter is usually positive (negative values mean the positive responses to the item are associated with lower values of the trait being measured, which usually means the item is keyed incorrectly). The magnitude of a indicates the strength of the relation between the item response and θ. The unit inside the exponential function in the denominator of equation (1), $a_i\theta + c_i$, is called the logit; that usually takes the form of linear regression (on θ) in IRT. The c parameter is the intercept of the linear regression in the logit; larger values of c mean more endorsement of the item. The algebraic form of the model in equation (1) is called the *slope-intercept* form; it is optimal for the computation involved in parameter estimation, but not so much for interpretation. For interpretation, the model is often written with the item parameters a_i and $b_i = -c_i/a_i$:

$$T_i(u_i = 1) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]}. \qquad (2)$$

Equation (2) is called the *slope-threshold* form, in which the parameter b_i is the value of the latent

variable θ at which a person has a 50–50 chance of endorsing the item. This is the feature of the model referred to earlier that "items have parameters that place them on the same scale as the variable being measured."

Following Birnbaum (1968), the logit of the 2PL model has sometimes been written $-1.7\tilde{a}_i(\theta - b_i)$; the constant 1.7 makes the value $\tilde{a} = a/1.7$ comparable to the discrimination parameter that would be obtained with a normal ogive model (Camilli, 1994). Because this historical reference has increasingly little value as the logistic replacement for the normal ogive has become ubiquitous, modern software like IRTPRO (Cai, Thissen, &, du Toit, 2011) reports all discrimination parameters as *a*. However, older software and research reports may report $\tilde{a}$, and further may refer to that value as *a*. Readers must rely on explicit statements of the IRT model being used, or a context that may report the use of specific software, to determine whether discrimination parameters reported in the literature are *a* or $\tilde{a}$.

To create a complete model for item response data, the population distribution for the construct being measured, the latent variable θ, must also be specified. Most often that distribution, φ(θ), is assumed to be $N(0,1)$, which specifies the shape of the population distribution (normal) and sets the otherwise-undefined units of scale of θ so that the mean is zero and the standard deviation is one. It is possible to estimate the shape of a nonnormal population distribution along with the item parameters; recently developed strategies to do that are described by Woods (2006, 2007), Woods and Thissen (2006), and Woods and Lin (2009).

From a statistical point of view, a crucial aspect of IRT models is that "the unobserved variable accounts for the observed interrelationships among the item responses." That means that, given (or *conditional* on) any particular value of the latent variable θ, the item responses are independent; this is usually called the assumption of *conditional independence* or *local independence*. Given that, the probability of a vector of item responses $\mathbf{u} = [u_1, u_2, \ldots u_I]$ for *I* items is

$$P(\mathbf{u}) = \int \left[\prod_i T_i(u_i) \right] \phi(\theta) d\theta. \qquad (3)$$

ML estimation of the item parameters involves choosing the values of a_i and c_i for all items, embedded in $T_i(u_i)$ in equation (3), such that the joint probability (likelihood) of all of the observed item responses is maximized.

The Rasch Model, and the One-Parameter Logistic Model

Rasch (1960) developed an item response model from a completely different set of principles than those used by Lazarsfeld, Lord, and Birnbaum in the evolution of the 2PL model. While Rasch (1960) used different notation, the trace line equation of the *Rasch model* can be written

$$T_i(u_i = 1) = \frac{1}{1 + \exp[-(\theta - b_i)]}. \qquad (4)$$

This form is like the 2PL model except that $a = 1$ (implicitly), and the mean and standard deviation of the latent variable are not specified in advance. Software designed to fit the Rasch model in the form of equation (4) usually sets the scale of the latent variable such that the average of the b_i parameters is 0; that, and the value $a = 1$, determines the scale in units of width called *logits*, with a population average of the latent variable that is some number of logits above or below the average of the b_is.

The Rasch model has only one parameter for each item, b_i, so it has sometimes been called the *one-parameter logistic* (1PL) model. However, to clarify nomenclature, Wainer, Bradlow, and Wang (2007, pp. 24–25) suggest the use of the term *Rasch model* for trace lines of the form of equation (4) and *1PL* for trace lines written as

$$T_i(u_i = 1) = \frac{1}{1 + \exp[-a(\theta - b_i)]}, \qquad (5)$$

in which a discrimination parameter *a* appears, but without a subscript, because it is equal for all items. Special estimation procedures are sometimes used to estimate the parameters of the Rasch model when the model is written as in equation (4). However, when the model is written as in equation (5), the same methods may be used to fit either the 1PL or 2PL models (or any of the other models described in this chapter), and the scale of the latent variable is set by the mean and variance of the population distribution φ(θ) (Thissen, 1982). When the variance of the population distribution is fixed at 1.0, the common discrimination parameter *a* must be estimated so that the slope of the trace lines is correct for that unit.

An Example with Dichotomous Responses: "Impulsivity"

THE DATA

To illustrate the way the 1PL and 2PL models fit data, and their use in item analysis, we consider the responses of 189 undergraduate students to

nine items of the Eysenck Personality Inventory Form A Extraversion scale (Eysenck & Eysenck, 1969). Revelle, Humphreys, Simon, and Gilliland (1980) divided the Extraversion scale into two subscales (impulsivity and sociability). The nine items forming the "impulsivity" subscale are shown in Table 18.1, along with the response (yes or no) that is keyed in the direction of impulsivity.

The item response data are from the Computer Administered Panel Survey (CAPS), sponsored by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. Each academic year between 1983–84 and 1987–88, approximately 100 undergraduates participated in that study, which involved responding to a large number of questionnaires via computer terminals. The Eysenck Personality Inventory was among the scales administered in 1987 and 1988. The data (for all of the questionnaires) remain publicly available at the Odum Institute's website; for this illustration, we use the data for the two academic years 1987 and 1988 (Odum Institute, 1988a).

THE 2PL MODEL

IRT item analysis for these data can begin with fitting the 2PL model—that is, the parameters *a* and *c* are estimated for each item. In this example, we used the implementation of the Bock-Aitkin (1981) EM algorithm in the computer program IRTPRO (Cai, Thissen, & du Toit, 2011) to compute the ML estimates of those parameters and to examine the goodness of fit of the model. The value of the M_2 goodness-of-fit statistic (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005, 2006) indicates some lack of fit ($M_2(27) = 47.45$, $p = .009$); however, the

Table 18.1 Items from a Nine-Item Impulsivity Subscale of the Eysenck Personality Inventory Form A

Item			*2PL*		*1PL*		*Special Model*					
#	Question	Key	a	s.e.	a	s.e.	a	s.e.	c	s.e.	b	s.e.
10	Would you do almost anything for a dare?	Y	0.83	0.42	0.83	0.10	0.98	0.12	−1.75	0.22	1.79	0.27
5	Do you stop and think things over before doing anything?	N	0.82	0.39	0.83	0.10	0.98	0.12	−1.67	0.21	1.71	0.27
8	Do you generally do and say things quickly without stopping to think?	Y	1.07	0.36	0.83	0.10	0.98	0.12	−0.42	0.17	0.43	0.18
22	When people shout at you, do you shout back?	Y	0.52	0.24	0.83	0.10	0.98	0.12	0.22	0.17	−0.22	0.17
39	Do you like doing things in which you have to act quickly?	Y	0.76	0.27	0.83	0.10	0.98	0.12	0.64	0.18	−0.65	0.19
13	Do you often do things on the spur of the moment?	Y	3.70	7.13	0.83	0.10	0.98	0.12	0.97	0.18	−1.00	0.21
3	Are you usually carefree?	Y	0.75	0.41	0.83	0.10	0.98	0.12	1.32	0.19	−1.35	0.23
1	Do you often long for excitement?	Y	1.18	0.37	0.83	0.10	0.98	0.12	1.87	0.22	−1.92	0.29
41	Are you slow and unhurried in the way you move?	N	0.13	0.19	0.83	0.10	0.11	0.20	0.55	0.15	−5.26	10.03
	−2loglikelihood		1918.13		1943.71		1931.41					
	M_2, *df*		47.45,	27	69.05,	35	58.26,	34				
	p, RMSEA		.009,	0.06	.001,	0.07	.006,	0.06				

Also given are the keyed response, estimated *a* parameters, their standard errors, and goodness-of-fit statistics from the 2PL and 1PL models and a special intermediate model; for the latter, the *c* and *b* parameters are also tabulated.

associated root mean square error of approximation (RMSEA) value (0.06) suggests this may be due to a limited amount of *model error*, and that the fit may fall in an acceptable range.

Goodness-of-fit statistics like M_2 are tests of exact fit, which follow their nominal distribution only if the model has no specification error. Such tests are generally unrealistic, because there must be some error in any strong parametric model (Browne & Cudeck, 1993). For IRT models, examples of trivial model misspecification that could induce significance in tests of exact fit include trace lines that are not *exactly* logistic, a population distribution that is not *exactly* normal, item parameters that are not *exactly* the same fixed values for all respondents, and small amounts of covariation among the item responses that arise from sources other than the latent variable being measured. Any or all of those elements of model misspecification may be present, yet the IRT model may provide a useful approximation to the data.

RMSEA is a measure that may be computed for any statistic like M_2; RMSEA provides a metric for model error. RMSEA was originally proposed by Steiger and Lind (1980); following suggestions by Browne and Cudeck (1993), it has become widely used in the context of structural equation modeling. Several rules of thumb have been suggested for values of RMSEA that suggest acceptable degrees of model error. Browne and Cudeck (1993, p. 144) suggested that "a value of the RMSEA of about 0.05 or less would indicate a close fit of the model … about 0.08 or less for the RMSEA would indicate a reasonable error of approximation … [one] would not want to employ a model with a RMSEA greater than 0.1." Hu and Bentler (1999) suggested that RMSEA values less than 0.05 or 0.06 may indicate acceptable fit. Chen, Curran, Bollen, Kirby, and Paxton (2008) point out that all of these rules of thumb for interpretation of RMSEA have been accompanied by cautions, and further that they perform unpredictably in practice. Nevertheless, Chen, Curran, Bollen, Kirby, and Paxton (2008, p. 491) conclude that "RMSEA and other fit indices have utility … These indices can supplement the chi-square in assessing the adequacy of a model in matching the data. However, sole reliance on a single fit index seems imprudent and we would recommend that multiple indices be examined." In the context of IRT, other fit indices include the $S\text{-}X^2$ item-level diagnostic statistics suggested by Orlando and Thissen (2000, 2003) and the $LD\ X^2$ statistics (Chen & Thissen, 1997) that will be discussed later in this chapter.

The *a* (slope, or discrimination) parameter estimates for the 2PL model are in Table 18.1 under the heading "2PL." The *a* parameters vary from 0.13 up to 3.7. However, we note that the corresponding standard errors vary from 0.19 up to 7.13; such large values are attributable to the small sample size (for IRT), less than 200 respondents. So before interpreting the variability among the slopes as indicative of reliable differences among the items' association with the latent variable being measured (impulsivity), we fit the 1PL model to the data to use the difference between the goodness of fit of the two models to compute a test of significance of the variation among the 2PL slopes.

THE 1PL MODEL

As shown in Table 18.1 under the heading "1PL," the ML estimate of the (single, common, equal) *a* parameter for all nine items is 0.83, with a standard error of 0.1; the standard error has become much smaller than the slope standard errors for the 2PL model because the data from all nine items are used to estimate the single common slope. The primary purpose of fitting the 1PL model is to obtain the value of −2loglikelihood to use with the corresponding value from the 2PL fit to test the significance of the variation among the *a* parameters in the 2PL model. For these data, the likelihood ratio test of the significance of variation among the slope parameters is computed as the difference between −2loglikelihood for the 1PL and 2PL models, 1943.71 – 1918.13 = 25.58, which is distributed as χ^2 on 8 *df*, $p = .0012$. The significance of that test statistic leads to the conclusion that there is some reliable difference in discrimination for these nine items; further data analysis can be used to identify that variation.

Table 18.2 tabulates the $S\text{-}X^2$ item-level diagnostic statistics suggested by Orlando and Thissen (2000, 2003) for the 2PL and 1PL fits to these data (as well as for a special intermediate model, to be described subsequently). Under the hypothesis of perfect model fit, these diagnostic statistics are approximately distributed as χ^2 values with the tabulated degrees of freedom; significant values indicate lack of fit. Because a very strong model such as the 2PL rarely fits perfectly, one expects some (slightly) significant values, because the model is not perfect. The statistics tabulated for the 2PL fit illustrate this point: four of the nine values are significant at the $p = .05$ level, but none has $p < .01$. For the 1PL model, however, the value of the $S\text{-}X^2$ for item 41 is very large: 25.41 on 5 *df*, $p < .0001$

Table 18.2 *S-X*² Item-Level Diagnostic Statistics for a Nine-Item Impulsivity Subscale of the Eysenck Personality Inventory Form A

	2PL			*1PL*			*Special Model*		
Item	X^2	*df*	*p*	X^2	*df*	*p*	X^2	*df*	*p*
1	7.66	5	.1757	5.93	5	.3145	6.23	5	.2860
3	10.66	5	.0584	10.53	5	.0614	10.67	5	.0582
8	11.29	5	.0458	12.54	6	.0509	11.34	5	.0450
10	3.59	5	.6111	3.83	5	.5747	3.43	5	.6348
13	6.63	3	.0844	16.27	5	.0061	13.97	5	.0157
22	9.74	6	.1359	11.41	5	.0437	13.11	5	.0223
39	14.92	5	.0107	14.94	5	.0106	14.94	5	.0106
5	12.15	5	.0327	11.70	5	.0390	12.85	5	.0247
41	13.29	5	.0208	**25.41**	5	**.0001**	13.23	5	.0213

Data are given for the 2PL and 1PL models and a special intermediate model.

(shown in bold in Table 18.2). That value suggests that closer inspection of the underlying frequency table is warranted.

Table 18.3 shows the underlying summed score by item response tabulation for the *S-X*² for item 41; italicized cells have fewer observed than expected; cells in roman type have more observed than expected. We notice that the reason for the large value of the *S-X*² diagnostic statistic for this item is that for low summed scores on the other items (0–4), we observe more "no" responses than expected, while for higher summed scores on the other items (5–8), we observe more "yes" responses than expected. This pattern suggests that the 1PL fitted value of the slope (*a*) parameter for item 41 is too high, producing expected values that are too high for low scores, and too low for high scores.

A SPECIAL MODEL WITH EQUALITY CONSTRAINTS

Having noted the large *S-X*² diagnostic statistic for item 41, the pattern of observed and expected values in Table 18.3, and also the fact that item 41 has by far the lowest estimated *a* parameter (in Table 18.1) for the 2PL fit, we are led to carefully examine the content of the item to see how it goes with the other eight items. Upon reflection, "Are you slow and unhurried in the way you move" does not appear to be a particularly cogent indicator of "impulsivity," as it might be reflected by responses to the other eight items. At this point, many item analysts would simply delete this item from the measure, on the grounds that it contributes almost no useful information, and it probably reflects other constructs. However, to illustrate the capabilities of IRT as a general model-fitting enterprise, here we refit the data with a special model that is a hybrid between the 1PL and 2PL models.

The special hybrid model imposes the constraint that the *a* parameters are equal for all of the items except item 41; however, it permits item 41 to have its own lower slope estimate. The item parameter estimates, standard errors, and goodness-of-fit statistics for this model are shown in the rightmost six columns of Table 18.1. Likelihood ratio tests indicate that this model fits significantly better than the 1PL model (G^2 = 1943.71 – 1931.41 = 12.3, 1 *df*, p = .0005), but it does not fit significantly worse than the 2PL model (G^2 = 1931.41 – 1918.13 = 12.3, 7 *df*, p = .066). The values of the M_2 overall goodness-of-fit statistic and its associated RMSEA are approximately the same for this hybrid model as they were for the 2PL model, although the special model estimates seven fewer parameters. The combined considerations of goodness of fit and parsimony suggest the use and interpretation of the special hybrid model.

Table 18.2 tabulates the *S-X*² item-level diagnostic statistics for the special model intermediate between the 1PL and 2PL in the rightmost columns; six of the nine values are significant at the p = .05 level. However, none are significant at the p = .01

Table 18.3 Underlying Summed Score by Item Response Tabulation for the $S\text{-}X^2$ Item-Level Diagnostic Statistic for the 1PL Fit to Item 41 of the Eysenck Personality Inventory Impulsivity Subscale

Response:	Yes		No	
Score	Observed	Expected	Observed	Expected
0				
1	*5*	*6.2*	5	3.8
2	*10*	*13.9*	16	12.1
3	*12*	*13.6*	18	16.4
4	*6*	*13.9*	31	23.1
5	20	11.2	*17*	*25.8*
6	9	6.7	*19*	*21.3*
7	7	3.7	*11*	*14.7*
8			3	2.6

For item 41 ($S\text{-}X^2(5) = 25.4$, $p = .0001$). Italicized cells have fewer observed than expected; cells set in roman type have more observed than expected.

Table 18.4 Underlying Summed Score by Item Response Tabulation for the $S\text{-}X^2$ Item-Level Diagnostic Statistics for the Special Hybrid Model Fit to Items 1 and 39 of the Eysenck Personality Inventory Impulsivity Subscale

Item 1	No		Yes	
Score	Observed	Expected	Observed	Expected
0	*1*	*1.3*	2	1.7
1	*2*	*2.9*	6	5.1
2	9	8.7	*22*	*22.3*
3	6	5.1	*18*	*18.9*
4	*6*	*7.7*	44	42.3
5	7	3.6	*25*	*28.4*
6	1	2.8	26	24.9
7			11	10.4
8			3	2.9
Item 39	**No**		**Yes**	
Score	Observed	Expected	Observed	Expected
0				
1	*5*	*7.2*	6	3.8
2	9	8.5	6	6.5
3	22	17.0	*14*	*19.0*
4	*10*	*16.8*	34	27.2
5	17	11.1	*20*	*25.9*
6	*6*	*7.0*	25	24.0
7	1	2.3	11	10.0
8			3	2.7

For item 1 ($S\text{-}X^2(5) = 6.2$, $p = .2860$); for item 39 ($S\text{-}X^2(5) = 14.9$, $p = .0106$). Italicized cells have fewer observed than expected; cells set in roman type have more observed than expected.

level, suggesting these statistics may indicate real, but negligible, misfit. Inspection of the underlying summed score by item response tabulations for these statistics confirms that. Table 18.4 shows the underlying summed score by item response tabulation for the $S\text{-}X^2$ item-level diagnostic statistics for the special hybrid model fit to items 1 and 39. The table for item 1 (in the upper panel of Table 18.4) shows perfectly good fit, with a nonsignificant value of $S\text{-}X^2$. The lower panel of Table 18.4 shows the summed score by item response table for item 39, which has a "significant" $S\text{-}X^2$ value of 14.9 on 5 *df* ($p = .0106$). We note that in Table 18.4, item 39, the worst-fitting of the nine items, has no particular tendency for higher observed than expected values to occur in blocks or "runs" that might indicate bad fit of the trace line, as we previously observed with item 41 fitted with the 1PL model (see Table 18.3).

To investigate the assumption of local independence, Chen and Thissen (1997) proposed the *LD* X^2 statistic, computed by comparing the observed and expected frequencies in each of the two-way cross-tabulations between responses to each item and each of the other items. These diagnostic statistics are (approximately) standardized χ^2 values (that is, they are approximately *z*-scores) that become large if a pair of items indicates local dependence (LD)—that is, if data for that item pair indicate a violation of the local independence assumption. Because the standardized *LD* X^2 statistic is only approximately standardized, and is based on a statistic with a long-tailed (χ^2) distribution, we do not consider values larger than 2 or 3 to be large. Rather, we consider values greater than 10 large, indicating likely LD; values in the range 5 to 10 lie in a gray area, and may indicate LD or may be a result of sparseness in the underlying table of frequencies. In practice data analysts use inspection of the item content, as

Table 18.5 Values of the Standardized LD X^2 Statistics for a Nine-Item Impulsivity Subscale of the Eysenck Personality Inventory Form A Fitted with the Special Model Intermediate Between the 1PL and 2PL Models

Item	1	3	8	10	13	22	39	5
3	−0.6							
8	−0.7	*−0.4*						
10	*−0.2*	*1.2*	−0.5					
13	*4.6*	*−0.7*	*2.6*	*−0.6*				
22	−0.5	2.7	*−0.1*	*−0.2*	−0.7			
39	−0.1	−0.7	0.6	−0.4	*−0.1*	*−0.2*		
5	6.6	*−0.6*	0.5	0.1	*−0.0*	2.2	*−0.6*	
41	*−0.6*	*−0.6*	1.3	*−0.2*	*−0.4*	0.0	*−0.5*	*−0.6*

Italic entries indicate positive LD, while roman entries indicate negative LD.

well as these statistics, to evaluate the presence of LD when it is indicated.

Table 18.5 shows the values of the standardized *LD* X^2 statistics for this impulsivity subscale fitted with the special model; italic entries indicate positive LD, while roman entries indicate negative LD. All of the values are relatively small, indicating no evidence of LD, and suggesting that the model fits satisfactorily.

The upper panel of Figure 18.1 shows the trace lines for the nine impulsivity subscale items. These curves plot the probability of the keyed "impulsive" response for each item as a function of the underlying latent variable. The curves show that the nine items are spread to cover the range of the impulsivity continuum; this is also apparent in the *b* parameters in Table 18.1, which range from 1.79 for item 10 (the rightmost curve in Fig. 18.1, and the first entry in Table 18.1 where the items are sorted by their *b* values) down to −1.92 for item 1 (the leftmost of the equal-slope curves in Fig. 18.1). The graphic also shows that for item 41, with its very low slope value (0.11), the probability of a "no" response changes very little across levels of impulsivity from lowest to highest. The *b* parameter estimate for item 41 is off-scale to the left, at −5.26, because the very low slope means the trace line does not descend to 0.5 until that point.

The center panel of Figure 18.1 augments the display of the trace lines from the upper panel with the information curves for each item. Information curves describe the additive contribution of each item to precision of measurement at each level of θ (Birnbaum, 1968). Each of the item information curves has a peak at the level of θ (impulsivity) that corresponds to its *b* value. The information curves are relatively low, reaching maximum values of only a little more than 0.2, because these items are not very discriminating.

In the lower panel of Figure 18.1, the information curves for each item are shown along with their sum (plus a constant 1.0, which is the information provided by the population distribution). The sum is the total information for the scale. Total information for this scale is approximately 2.0 for a wide range of the continuum. That means that the standard errors of IRT scores computed for this scale are approximately $\frac{1}{\sqrt{2}} = 0.7$. It also means that an IRT approximation of reliability is approximately 0.5, with "reliability" computed as one minus measurement variance.

CONCLUSION ON IMPULSIVITY

If we were constructing this scale based on the IRT analyses, we would omit item 41. As shown in Figure 18.1, it provides negligible information; the low slope implies that the responses to that item are not very related to the construct of impulsivity as defined by the other eight items. The remaining eight-item scale appears to provide good, if not terribly precise, measurement of impulsivity. Measurement precision is approximately constant across a wide range of the construct.

The relatively uniform information shown in the lower panel of Figure 18.1 is rarely obtained with measures of pathology. For the general

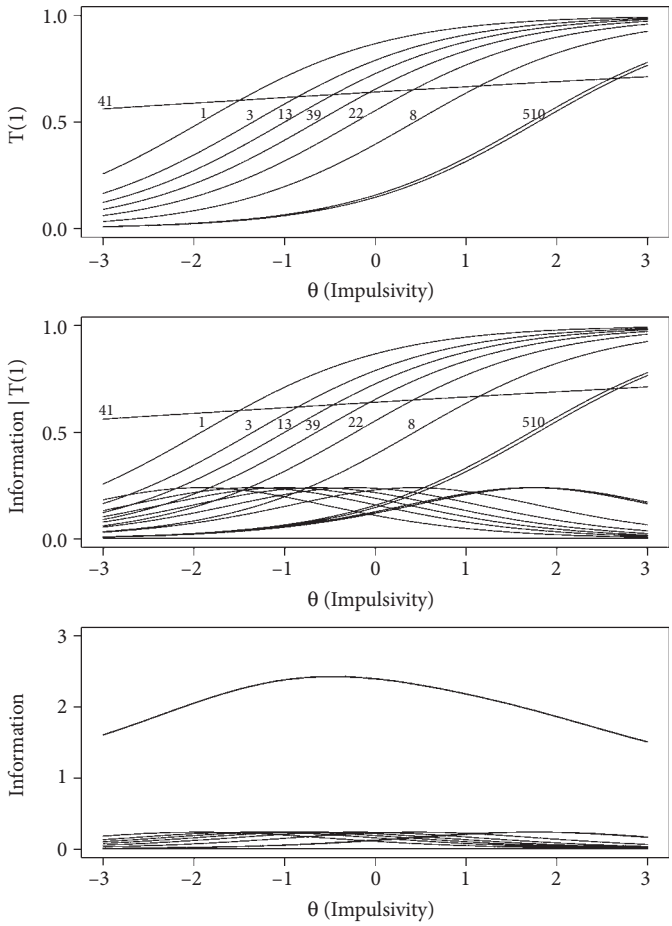


Figure 18.1 Upper panel: trace lines for the nine impulsivity subscale items. Center panel: trace lines from the upper panel augmented with the information curves for each item. Lower panel: the information curves for each item with their sum (plus a constant 1.0 from the population distribution), which is the total information for the scale.

population, responses to items measuring clinical symptoms produce test information curves that show peaked information for a range of the continuum over which the items discriminate among persons, and much less information elsewhere. For examples, scales measuring anxiety or depressive symptoms in the general population may have high information only for the top half or quarter of the continuum.

Scale Scores

Scale Scores for Response Patterns

IRT scale scores are statistical estimates of the level of the latent variable (θ) associated with the respondent's item responses. Scale scores are based on a function of the item responses that is usually called the *posterior*; that is

$$\Box(\mathbf{u}) = \left[\prod_i T_i(u_i)\right]\phi(\theta), \qquad (6)$$

which is the product of the trace lines for the particular item responses in the response pattern and the population distribution φ(θ). The term *posterior* is borrowed from Bayesian statistics, in which a posterior is the product of a *prior* density (here, φ(θ))and a *likelihood* (here, the product of the trace lines).

Figure 18.2 provides graphical illustrations of the components of equation (6) for two response patterns to the nine-item "impulsivity" scale discussed earlier in this chapter, using the item parameters from the "special" model. The upper panels of both the left and right sides of Figure 18.2 show the $N(0,1)$ population density φ(θ). The middle panel of the left side of Figure 18.2 shows the nine trace lines associated with the response pattern **u** = 000001110 for the items in the order listed in Table 18.1; this response pattern has a summed score of 3, but that is not the IRT scale score. The scale score is an estimate of the center of the

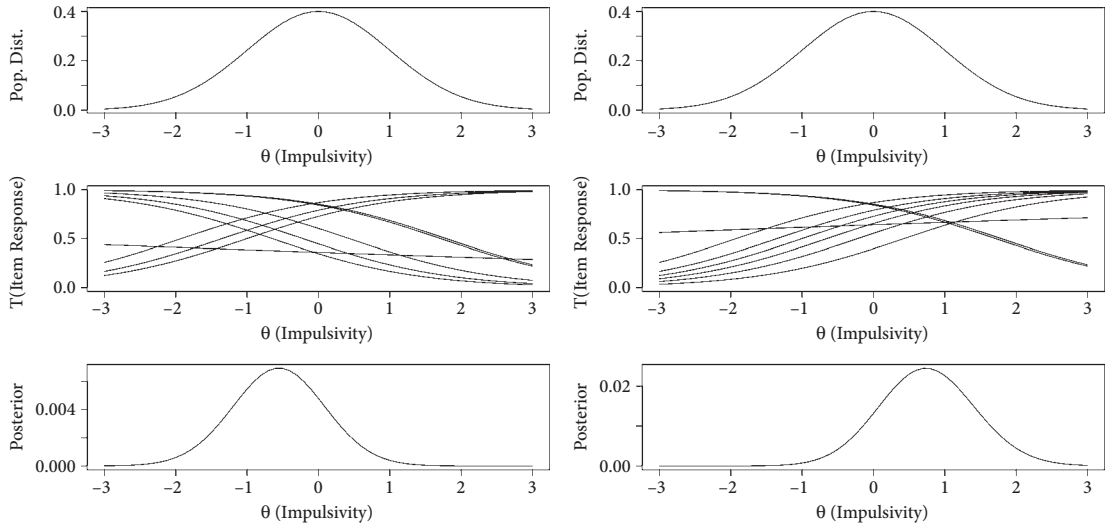


Figure 18.2 Upper panels, left and right: the $N(0,1)$ population density $\phi(\theta)$. Center panels: trace lines associated with the response pattern $u = 000001110$; right, for response pattern $u = 001111111$. Lower panels: the posterior densities for $u = 000001110$, left, and $u = 001111111$, right.

posterior shown in the lower panel of the left side of Figure 18.2; that posterior is the product of the ten curves in the two panels above—the population density and the nine trace lines for 000001110.

Either of two estimates of the center of the posterior, the mean or the mode, is commonly used as an IRT scale score. The mean of the posterior density in the lower-left panel of Figure 18.2 is −0.56, which is a scale score 0.56 standard deviation units below average. (IRT scale scores are usually computed in z-score units.) The mean is usually referred to as the *expected a posteriori* (EAP) value (Bock & Mislevy, 1982), and the corresponding posterior standard deviation is usually reported as the standard error of the scale score (0.65 for this example). A computational alternative to the EAP is the *maximum a posteriori* (MAP) estimate, which is the mode of the posterior (−0.55 for this example). An estimate of the standard deviation of the posterior derived from the degree of curvature of the density at the mode is reported as the standard error of MAP scale scores (0.64 for this example). The EAP and MAP, and their corresponding standard error values, are not exactly the same because the posterior densities are not perfectly symmetrical; however, they are usually very similar, as they are in this case.

The right panel of Figure 18.2 shows the same components for a second response pattern, $\mathbf{u} = 001111111$ for the items in the order listed in Table 18.1. This response pattern is associated with higher scale score values: the EAP is 0.76, with a standard deviation of 0.67, and the MAP is 0.73, with a standard error of 0.67.

Scale Scores for Summed Scores

IRT scale scores can be associated with summed scores, but their computation is not as simple as summation. Thissen and Orlando (2001) and Thissen, Nelson, Rosa, and McLeod (2001) describe the use of EAPs associated with the IRT posterior for summed scores for scales using dichotomous and polytomous items, respectively. The IRT posterior for a summed score x is the sum of the posteriors for all of the response patterns with that summed score,

$$\square(x) = \sum_{\sum u_i = x} \left[\prod_i T_i(u_i) \right] \phi(\theta). \qquad (7)$$

For scales with many items and/or item response categories, brute-force computation of the summed-score posterior in equation (7) is not feasible. However, Thissen, Pommerich, Billeaud, and Williams (1995) described the use of a recursive algorithm that may be used to compute the posterior in equation (7), and then its EAP value and the corresponding standard deviation.

Table 18.6 shows the values of the EAPs and corresponding posterior standard deviations for the ten summed scores 0 through 9 for the nine-item "impulsivity" scale, again using the parameters of the "special" model. From the point of view of scale score computation, a side effect of the use of the

Table 18.6 Summed Score to Scale Score Conversion Table for the Nine "Impulsivity" Items

Summed Score	EAP[θ|x]	SD[θ|x]	Modeled Proportion	Observed Proportion
0	−1.86	0.70	0.01	0.01
1	−1.52	0.70	0.03	0.02
2	−1.14	0.69	0.07	0.08
3	−0.74	0.68	0.13	0.15
4	−0.35	0.68	0.18	0.13
5	0.04	0.68	0.20	0.27
6	0.44	0.69	0.19	0.14
7	0.84	0.70	0.13	0.14
8	1.26	0.71	0.06	0.06
9	1.70	0.72	0.01	0.02

recursive algorithm to compute the scores is that it also computes a model-based estimate of the proportion of respondents with each summed score; those values are also in Table 18.6, along with the observed summed-score distribution. Comparing the two columns on the right of Table 18.6 is a way to see the degree to which the IRT model reproduces the summed-score distribution. Lord and Wingersky (1984) originally used the recursive algorithm to compute the model-implied summed-score distributions for dichotomous items; Thissen and colleagues (1995) described the generalization for polytomous items and its use to compute scale scores.

Scale scores for response patterns and for summed scores each have advantages and disadvantages. Advantages of response pattern scores include greater precision, because they use all available information (this advantage may be small), and the fact that they can be used to score any arbitrary collection of items. The most salient disadvantage of response pattern scale scores is computational complexity—special-purpose software is required. Another feature of response pattern scale scores that may be a disadvantage in educational measurement is that respondents with the same summed score may have different response pattern scores.

The primary advantage of scale scores for summed scores is the simplicity of their use after they are once tabulated in a score-translation table like Table 18.6: the scale scores may be assigned to respondents with a simple score-substitution algorithm, using the values in the table. The chief disadvantage of the summed-score scale scores is that the posteriors for summed scores have larger standard deviations, unless Rasch-family models are used; they provide (slightly) less precise measurement.

To illustrate the small differences between scale scores for response patterns and their corresponding summed scores, recall that the EAP and its corresponding standard deviation for the response pattern $\mathbf{u} = 000001110$ in the left side of Figure 18.2 are −0.56 and 0.65. From Table 18.6, we see that the EAP for a summed score of 3 is −0.74 with a posterior standard deviation of 0.68. The response pattern and summed score EAPs differ because the summed-score estimate includes all posteriors that have a summed score of 3; there are different IRT scale scores (EAPs) associated with each of those response patterns, because the response pattern score (effectively) weights each item response by the item's discrimination. The posterior standard deviation for the summed-score posterior is slightly larger (0.68 vs. 0.65) due to this aggregation. For the example in the right side of Figure 18.2, the EAP and its corresponding standard deviation for response pattern $\mathbf{u} = 001111111$ are 0.76 and 0.67. From Table 18.6, we see that the EAP for a summed score of 7 is 0.84 with a posterior standard deviation of 0.70.

We have illustrated summed-score EAPs with binary response data; however, we note that such scoring generalizes to polytomous response items.

Models for Items with Polytomous Responses

The Graded Model

Samejima's model (1969, 1997) for graded item responses is often applied to the analysis of items that are accompanied by Likert-type response scales. The model for an item with K ordered response alternatives $k = 0, 1, 2, \ldots K-1$ is

$$T_i(u_i = k) = T_i^*(u_i = k) - T_i^*(u_i = k+1) \qquad (8)$$

in which $T_i^*(u_i = 0) = 1$ and $T_i^*(u_i = K) = 0$.

$$T_i(u_i = k) = \frac{1}{1+\exp[-a_i\theta + c_{ik}]} - \frac{1}{1+\exp[-a_i\theta + c_{ik+1}]} \qquad (9)$$

$$T_i(u_i = k) = \frac{1}{1+\exp[-a_i(\theta - b_{ik})]} - \frac{1}{1+\exp[-a_i(\theta - b_{ik+1})]} \qquad (10)$$

The graded model divides the responses into binary pieces; each T_i^* is a 2PL model for the probability of a response in a category or higher (e.g., probability of a response in category 2 or higher, probability of a response in category 3 or higher, etc.). Equations (9) and (10) are the graded model in slope-intercept and slope-threshold form, respectively. Our description will focus on the slope-threshold form parameterization. The a parameter is the slope or discrimination parameter, and the b_k parameter is the threshold for a category or higher. The value of b_k is the point on the construct axis at which the probability that the response is in category k or higher passes 50 percent. $T_i^*(u = k)$ is the trace line describing the probability that a response is in category k or higher, for each value of the underlying construct. The probability that a response is in a particular category (k) is the probability of observing category k or higher minus the probability that the response is in category $k + 1$ or higher.

An Example with Polytomous Responses: "Impulsiveness"

THE DATA

To illustrate Samejima's model for graded item responses and its use in item analysis, item response data for five items from the Barratt Impulsiveness Scale (BIS) will be used. The BIS is a 30-item, commonly used measure of the personality construct of impulsiveness for research and clinical settings (Stanford, Mathias, Dougherty, Lake, Anderson, & Patton, 2009). The item response data were obtained from 1,178 undergraduates at Baylor University; we thank Matt Stanford for the use of these data. Table 18.7 lists the content for the five items; responses were made on a 4-point Likert-type scale (1 = rarely/never, 2 = occasionally, 3 = often, 4 = almost always/always).

THE GRADED MODEL

Table 18.7 lists the item parameters for the five BIS items. The M_2 goodness-of-fit statistic, especially as an RMSEA value, suggests reasonably good fit ($M_2(85) = 144.68$, $p < .001$; RMSEA = 0.02). In the analysis, the responses 1 through 4 correspond to response categories 0 through 3, respectively, in the graded model; the response labels 1 through 4 are used in the graphics. Figure 18.3 shows the full workings of the graded model. The upper panel

Table 18.7 Graded Model Parameter Estimates for Five BIS Items

Item		a	s.e.	b_1	s.e.	b_2	s.e.	b_3	s.e.
BIS2	I do things without thinking	2.41	0.23	−0.56	0.05	1.33	0.07	2.63	0.15
BIS5	I don't pay attention	1.02	0.09	−0.90	0.09	1.79	0.14	3.93	0.32
BIS8	I am self controlled	0.88	0.08	−0.52	0.09	2.12	0.19	4.77	0.45
BIS14	I say things without thinking	1.55	0.12	−0.94	0.07	1.23	0.08	2.62	0.16
BIS19	I act on the spur of the moment	1.36	0.10	−1.56	0.10	0.97	0.08	2.67	0.17

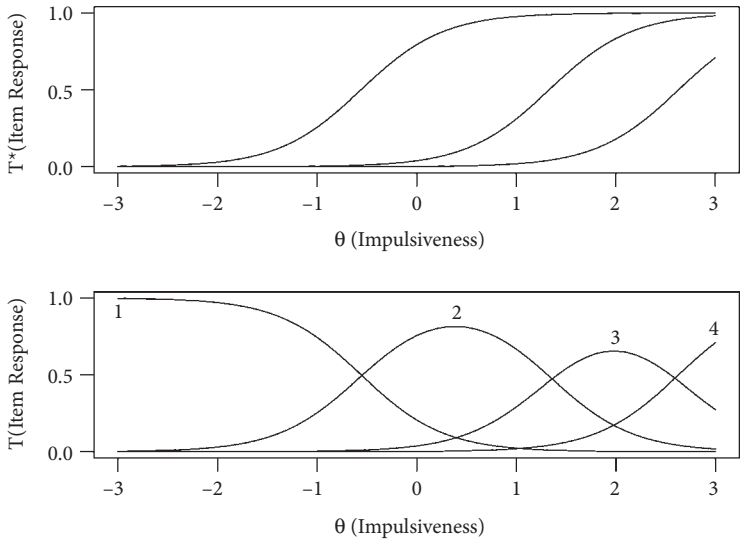


Figure 18.3 Upper panel: trace lines for $T^*(u = k)$. Lower panel: trace lines for the probability that a response is in a particular category, $T(u_i = k)$.

shows the trace lines for the individual binary (2PL) pieces, $T_i^*(u = k)$, representing the probability that a response is in category 1 or higher, category 2 or higher, or category 3. Because a single slope parameter is estimated for each item, the trace lines are horizontally offset identical ogives. For the leftmost trace line, $b_1 = -0.56$; for the rightmost trace line, $b_3 = 2.63$. Higher threshold values imply that greater amounts of the trait are required to observe a response in a category or higher.

The lower panel shows the trace lines for the probability that a response is in a particular category, $T_i(u_i = k)$, as a function of the value of the underlying construct. The slope parameter for this item is 2.41; that quantifies the strength of the relation between the item response and the underlying construct as defined by the other items in the analysis. The threshold parameters quantify the ease or difficulty of endorsement of each response category and are reflected in the left–right locations and the heights of the trace lines. For example, the probability of endorsing a "1" decreases as the value of the underlying construct increases; the probability of endorsing a "2" is most likely between values of the construct between −0.5 and +1.

Figure 18.4 shows the trace lines for the probability of observing each categorical response for each of the five BIS impulsiveness items as a function of the value of the construct. Inspection of the trace lines for the five items shows differences in the magnitude of the slope parameters. For example, the trace lines for the item "I do things without thinking" shows considerable discrimination among the response options compared to the trace lines for the item "I am self controlled." The slope parameter is highest for "I do things without thinking" and lowest for "I am self controlled." For "I am self controlled," trace lines show that the probability of observing option "2" (occasionally) spans the construct continuum and overlaps with the probability of observing options "3" (often) and "4" (almost always). The low slope for "I am self controlled" implies that the responses to this item are less related to the underlying construct as defined by the other four items. Possibly, the fact that "I am self controlled" is reverse-scored because it contraindicates the trait of impulsiveness contributes to the lower relation of the item response to the construct.

The right–left shifts among the item trace lines for each item show the differences in ease of endorsement of the response alternatives for the five items. For example, the response alternatives 1 = rarely, 2 = occasionally, and 3 = often are easier to endorse (i.e., have lower threshold parameters) for the item "I act on the spur of the moment" compared to the item "I do things without thinking;" there is little difference in the threshold parameters for the response option 4 = almost always between these two items.

CONCLUSION ON IMPULSIVENESS

The data examined in this example illustrate "textbook" item performance and good fit of the graded model. They do so because this five-item set

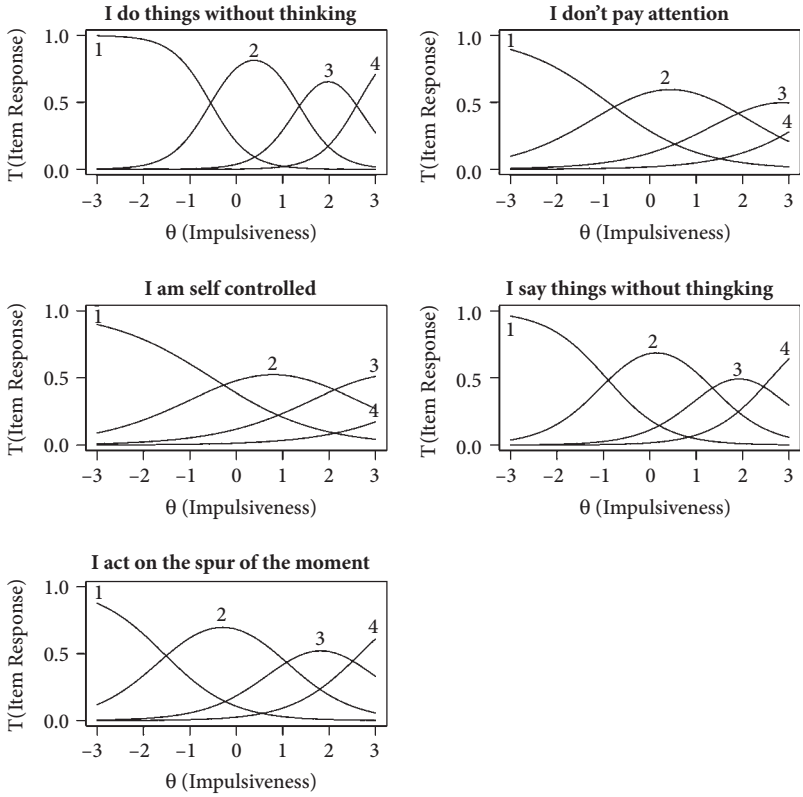


Figure 18.4 Trace lines for the probability of observing each categorical response for each of the five BIS impulsiveness items.

was selected for that purpose. Later in this chapter, in the section on multidimensional models, we will revisit these items in a more realistic context.

The Nominal Model

Another model that is sometimes used for items with polytomous responses is the nominal categories model (Bock, 1972, 1997b; Thissen, Cai, & Bock, 2010). Unlike the graded model, the nominal model does not require the order of response categories to be specified in advance; the analysis indicates how the response categories are mapped onto the latent variable. It does so at the expense of estimating additional parameters, which requires more data. It is also more complex to use and interpret than the graded model, so it is not so often used in its general form.

A constrained version of the nominal model is called the *generalized partial credit* (GPC) model (Muraki, 1992, 1997). The GPC model imposes restrictions on the parameters of the nominal model requiring the response categories to be ordered (Thissen, Cai, & Bock, 2010; Thissen & Steinberg, 1986); it uses the same number of parameters as the graded model and produces similar (but not identical) trace lines.

The nominal model may be used for two purposes. The first is to check putatively ordered responses to see if the data support the assumptions of the GPC or graded models. A second use of the nominal model is for item responses that are polytomous, but are not uniformly ordered. This may be the case for items that use response alternatives that combine "never" with a frequency scale (see for an example Revicki, Chen, Harnam, Cook, Amtmann, Callahan, Jensen, & Keefe, 2009) or when response pattern testlets are constructed from two or three dichotomous items. Steinberg and Thissen (1996) describe the use of the nominal model to fit testlets that are combinations of dichotomous items that are otherwise locally dependent as individual items.

Differential Item Functioning

Differential Item Functioning in Items with Dichotomous Responses

Differential item functioning (DIF) means that an item response is associated differently with the latent variable being measured for one subgroup of a population than another. An item exhibiting DIF must be less valid for at least one of the groups involved, because it is indicating that something else,

in addition to the construct being measured, influences item responses. Setting aside items exhibiting DIF may increase the validity of the test. DIF analysis originated as a solution to challenges to validity in educational measurement (Holland &Wainer, 1993); however, it has increasingly come to be used as a tool for measurement of personality and psychopathology, and in experimental social psychology (examples include reports by Collins, Raju, & Edwards, 2000; Ellis, 1989; Hancock, 1999; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Orlando & Marshall, 2002; Reeve, 2000; Schaeffer, 1988; and Steinberg, 1994, 2001).

The language of IRT is particularly well suited to the definition of DIF; a succinct statement of the meaning and consequences of DIF was provided by Lord (1980, p. 212): "If … an item has a different item response function for one group than for another, it is clear that the item is biased." Because item response functions (trace lines) are in one-to-one correspondence with (sets of) item parameters, the statistical detection of DIF involves some test of the null hypothesis that an item has the same parameters for both groups. Such statistical tests are straightforward, and their mechanics have been described by Thissen, Steinberg, and Wainer (1988, 1993).

The upper panel of Figure 18.5 shows trace lines for two groups (men and women, from the nonclinical standardization sample) for endorsement of the item "I cry easily" on the Minnesota Multiphasic

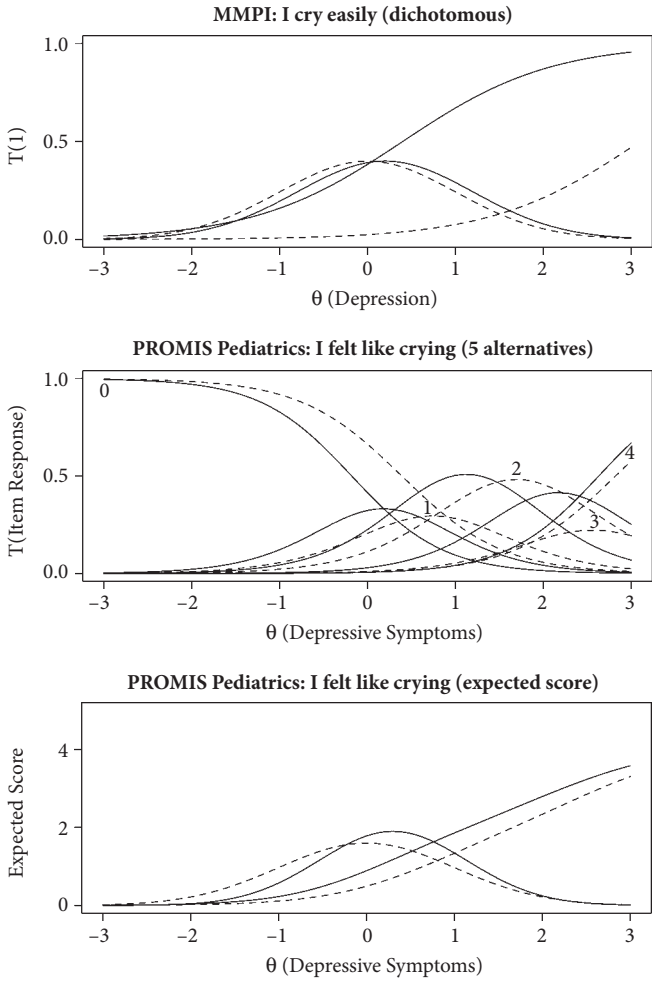


Figure 18.5 Upper panel: Trace lines for two groups (women, solid line; men, dashed line) for endorsement of the item "I cry easily" on the MMPI-2; the normal curves are the population distributions for women (solid) and men (dashed). Center panel: Trace lines for the item "I felt like crying" (0 = never, 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always) for girls (solid lines) and boys (dashed lines) from the item tryout for the PROMIS Pediatric Depressive Symptoms Scale. Lower panel: Expected score curves for the item "I felt like crying" computed from the trace lines in the center panel for girls (solid line) and boys (dashed line).

Personality Inventory (MMPI-2) (Graham, 2000), calibrated in the context of the items making up the Harris-Lingoes Subjective Depression Subscale (D1). This analysis was one of many reported by Reeve (2000). The two trace lines shown in the upper panel of Figure 18.5 are not identical: the curve for the male respondents is shifted well to the right of the curve for the female respondents, indicating that, for men, much higher levels of depression are required before endorsement of the item "I cry easily" becomes likely. This difference is expressed numerically in the threshold parameters (*b*): for women $b = 0.4$ for this item, while for men $b = 3.1$.

IRT's separation of the concepts of the trace lines and the population distribution is used to make a crucial distinction in DIF analysis. At the level of responses to an individual item, for example, if we observe only that more women than men endorse the item "I cry easily," we do not know if that is attributable to a sex-related difference in response to that particular item, or because women are more depressed on average than men. However, in the context of the analysis of a set of items, IRT can be used to estimate the parameters of the population distribution from a set of "anchor" items assumed to function the same for both groups, and then perform a test of the difference between the *studied* or *candidate* item's parameters for the two groups. The resulting test of DIF is conditional on, or corrected for, the overall difference in the distributions of the latent variable between the two groups, as defined by the anchor item set. In the upper panel of Figure 18.5, the estimated population distributions for depression for men and women are shown as the normal distributions. The dashed curve is for men, who are the *reference group* in this analysis, so their population distribution is normal with a mean of zero and a variance of one. Using all of the other items on the Harris-Lingoes Subjective Depression Subscale of the MMPI-2 as the anchor set, the population distribution for women (the *focal group*) is estimated to have a mean 0.2 standard units higher on depression than the average for the men; that normal curve is solid in Figure 18.5.

The upper panel of Figure 18.5 illustrates *uniform* (Mellenbergh, 1982) DIF: the probability of endorsement of the item is uniformly higher for women, across the entire range of the latent variable. In other items or analyses, there can also be *nonuniform* DIF, in which one group is more likely to endorse the item over part of the range of θ, and then the trace lines cross and the other group becomes more likely to respond positively. For logistic models, uniform DIF corresponds to differences in the intercept (*c*) or threshold (*b*) parameters, and nonuniform DIF corresponds to differences in the discrimination (*a*) parameters.

DIF in Items with Polytomous Responses

DIF analysis based on IRT generalizes straightforwardly to items with polytomous responses: DIF still means that the trace lines differ between groups. The center panel of Figure 18.5 shows the graded model trace lines for the item "I felt like crying" fitted to five response alternatives (0 = never, 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always) for boys and girls. This item was among the items considered for inclusion in the Patient-Reported Outcomes Measurement Information System (PROMIS) Pediatric Depressive Symptoms Scale (Irwin, Stucky, Thissen, DeWitt, Lai, Yeatts, Varni, & DeWalt, 2010). It was calibrated with the other depressive symptoms items on the same item tryout form, and checked for DIF; because the item exhibited DIF, it was set aside and not included in the final item pool.

While it could be said that the center panel of Figure 18.5 shows that the girls' (solid) trace lines are shifted to the left relative to the boys' (dashed) trace lines, it could also be said that it is difficult to interpret a graphic showing two sets of five curves. The *expected score curves* shown in the lower panel of Figure 18.5 summarize the DIF more clearly. The expected score curve for an item fitted with a polytomous IRT model is the expected value, or average, of the item responses (with their categorical index numbers 0, 1, 2, 3, and 4 taken to have their numerical value), so the curve rises from an expected value of zero to four. In the lower panel of Figure 18.5, the girls' (solid) line is about 0.67 points higher than the boys' (dashed) curve for much of the range, meaning we expect girls to score almost a point higher on this item than boys at the same level of underlying depressive symptoms. As in the dichotomous example, this analysis separates differences in level of depressive symptoms overall for boys and girls from the DIF. The population distribution for girls is estimated to have a mean 0.3 standard units higher on depression than the average for the boys, with a slightly smaller standard deviation (0.8 for girls vs. the reference 1.0 for boys); those normal population distributions are the solid and dashed (respectively) bell-shaped curves in the lower panel of Figure 18.5.

The reader has probably noticed the similarity (except for the amount of DIF) between the upper

and lower panels of Figure 18.5. Both illustrate the sex-related DIF associated with items involving "crying" on depression scales. "Crying" items appear to invariably exhibit sex-related DIF on depression scales. In addition to the two examples illustrated here, other published accounts of items involving "crying" on depression scales that exhibit DIF between gender groups include Schaeffer's (1988) analysis of the Hopkins Symptom Checklist; Santor, Ramsay, and Zuroff's (1994) analyses of the Beck Depression Inventory; Reeve's (2000, 2003) analysis of the clinical standardization sample for the Harris-Lingoes Subjective Depression Subscale of the MMPI-2; analyses of the CES-D by Cole, Kawachi, Maller, and Berkman (2000), Gelin and Zumbo (2003), Yang and Jones (2007), and Covic, Pallant, Conaghan, and Tennant (2007); and the analysis of the PROMIS adult depression scale by Teresi, Ocepek-Welikson, Kleinman, Eimicke, Crane, Jones, and colleagues (2009).

"Crying" items on depression scales provide easily comprehensible illustrations of the meaning of DIF, and the reason that items that exhibit DIF are often set aside in scale construction, even when such items appear at first glance to be indicators of the trait being measured. Reference to the concepts of *sensitivity* and *specificity* is useful: crying is a recognized symptom of depression, so questions about crying are *sensitive* indicators of depression. However, crying is not specifically associated with depression; there are many other situational contexts and individual difference variables that also lead to crying, so questions about crying lack specificity. Put simply, they measure or indicate other variables as well as depression. It happens that some of these other differences are sex-related, so that DIF analyses between gender groups yield significant differences for crying items. If items involving crying are included in a depression scale, female respondents tend to score somewhat higher than male respondents at the same level of depression, just because the scale includes the crying item. Thus, crying items are set aside from contemporary depression scales, like those constructed by the PROMIS teams using IRT.

Items on many other topics, embedded in scales measuring many other constructs, may also exhibit DIF with respect to many other grouping variables. Sometimes the source of the DIF is as easy to understand as crying items on depression scales; sometimes it remains mysterious. Nevertheless, when items exhibit substantial DIF, it is common practice to exclude them from the scale. DIF analysis can also be used as a method to enhance understanding of the construct, as in the analysis of the DSM-IV criteria for major depression by Carragher, Mewton, Slade, and Teesson (2011).

An Example: A Randomized Groups Experimental Analysis of Item Positioning

While the origins of DIF analysis lie in the observational comparison of responses from preexisting (usually demographically defined) groups, DIF procedures can also be used to analyze item response data collected in randomized experiments. Examples include the investigation of serial position or context effects in personality measurement (Steinberg, 1994, 2001).

THE DATA

The questionnaire considered in this example included 12 extraversion items modified from the International Personality Item Pool (IPIP) (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006) listed in Table 18.8. Many personality instruments, like the IPIP Extraversion items, include items that are trait-indicating (i.e., extraversion) and items that contraindicate the construct (i.e., introversion). The original intent of the study was to investigate how respondents develop a sense of what the questionnaire is measuring as they respond to trait-indicating and trait-contraindicating items. The items were arranged in one of three different orders: the extraversion and introversion items were either blocked or mixed as shown by their serial position numbers in the "Form" columns in Table 18.8. Participants responded to one of the forms and were asked to indicate "how much each statement describes you" on a 5-point Likert-type response scale. The numbers 1 through 5 were provided as response alternatives, and the end points were labeled "Not at all" and "Very much." The first seven items (labeled "E" in Table 18.8) are positive statements indicating extraversion; the final five items are positive statements indicating introversion and are labeled "I" in the table. For fitting the graded model, all items were scored so that the highest response category indicates more extraversion; that is, the "I" items in Table 18.8 were reverse-scored.

The item response data were collected in classrooms at the University of Houston. The three forms of the questionnaire were randomly assigned to participants. The items were presented in the order shown in Table 18.8 on Form 2, with the "E" items first; on Form 3 the items were reordered so

Table 18.8 The Items and Their Serial Positions on the Three Forms of the Randomized Groups DIF Study

Form 1 Mixed	Form 2 "E" first	Form 3 "I" first	Extraversion/ Introversion	
1	1	6	E	I feel comfortable around people.
4	2	7	E	I start conversations.
5	3	8	E	I talk to a lot of different people at parties.
7	4	9	E	I make friends easily.
10	5	10	E	I feel at ease with people.
11	6	11	E	I am skilled at handling social situations.
12	7	12	E	I am the life of the party.
8	8	1	I	I am a very private person.
9	9	2	I	I often feel uncomfortable around others.
2	10	3	I	I keep in the background.
3	11	4	I	I am quiet around strangers.
6	12	5	I	I find it difficult to approach others.

that the "I" items were presented first. Form 1 interleaved "I" and "E" items in the order shown in the first column of Table 18.8. The sample sizes for the three forms were $N = 347$, 359, and 353 for Forms 1, 2, and 3, respectively.

In place of focusing the analysis on the original blocked versus mixed order of the items, for this illustration these data are used to examine a simpler question—that is, the degree to which item responses vary as items are repositioned on a scale. If the items can be repositioned without changing what they indicate about the respondents' introversion–extraversion, then IRT analyses should yield the same item parameters for each item (within sampling error) across the three groups. If there are context or position effects, the item parameters for any affected item would differ across these groups to whom the items were presented in different orders—and that is what DIF analysis detects.

Because the groups are defined by the random assignment of the forms, the population distribution of introversion–extraversion is known to be the same for all three groups. So in the DIF-like analyses that follow, we assume that the latent variable is distributed $N(0,1)$ within each of the three groups—unlike "demographic" DIF analysis, no groups' population means or variances are estimated. There is no "anchor" set of items used to estimate the population means and variances.

Preliminary analyses of the data indicated that the "E" and "I" items may not be represented well with a single unidimensional model. Therefore, the DIF analyses that follow consider the "E" and "I" items separately, although the existence and positions of the items in the "other set" provide experimentally manipulated context for each.

DIF analysis, "E" items

Table 18.9 shows the DIF statistics for the analysis of the seven "E" items. With three groups, for each item there are two comparisons, or contrasts, just as in the analysis of variance. In this case the first contrast compares the trace lines for Forms 1 and 2 (averaged) with those for Form 3, and the second contrast compares the results from Forms 1 and 2. Thus, there are 14 overall significance tests for DIF in Table 18.9; the only one that provides evidence of DIF, after controlling the false discovery rate using the Benjamini and Hochberg (1995) (B-H) procedure, is for item 1–1-6 (the first item on Forms 1 and 2, and the sixth item on Form 3),

Table 18.9 DIF Statistics for the Analysis of the Seven "E" Items

Item numbers in:												
Group 1	Group 2	Group 3	Contrast	Total X^2	*df*	*p*	X^2_a	*df*	*p*	$X^2_{c\|a}$	*df*	*p*
1	1	6	1	23.1	5	.0003	0.6	1	.4449	22.5	4	.0002
			2	10.7	5	.0565	0.4	1	.5388	10.4	4	.0346
4	2	7	1	7.3	5	.1957	0.2	1	.6792	7.2	4	.1267
			2	6.2	5	.2854	5.0	1	.0258	1.3	4	.8672
5	3	3	1	4.7	5	.4606	0.0	1	.8564	4.6	4	.3296
			2	3.3	5	.6579	0.1	1	.7906	3.2	4	.5246
7	4	9	1	2.9	5	.7114	0.4	1	.5128	2.5	4	.6452
			2	13.3	5	.0207	4.7	1	.0308	8.6	4	.0708
10	5	10	1	3.5	5	.6255	0.2	1	.6890	3.3	4	.5050
			2	10.6	5	.0604	5.8	1	.0165	4.8	4	.3071
11	6	1	1	3.9	5	.5585	0.3	1	.5634	3.6	4	.4626
			2	3.1	5	.6864	1.8	1	.1779	1.3	4	.8667
12	7	12	1	1.3	5	.9314	0.5	1	.4677	0.8	4	.9378
			2	8.6	5	.1263	0.0	1	.8571	8.6	4	.0730

"I feel comfortable around people." When that is the first item, it has a different set of trace lines than when it is preceded by the five "I" items. Table 18.10 shows the parameter estimates for the three groups for this item.

Figure 18.6 shows the differences between the fitted trace lines (upper panel) and expected score curves (lower panel) for item 1–1-6; the dashed lines are for the average of the item parameters for Forms 1 and 2, while the solid lines are for the significantly different item parameters for Form 3. When this item is preceded by the five "I" items (the solid lines in Figure 18.6), higher levels of θ (Extraversion) are required to endorse responses 2, 3, 4, or 5 than is the case when "I feel comfortable around people" is the first item the respondent answers.

DIF ANALYSIS, "I" ITEMS

Table 18.11 shows the DIF statistics for the analysis of the five "I" items. This analysis used the same two contrasts among the three groups as were used in the analysis of the "E" items. The largest three X^2 values are significant after using the B-H procedure to control the false discovery rate; as illustrations we tabulate and discuss the effects associated with the largest two of those. The effect for item 8–8-1 is large for the first contrast, between the two forms on which "I am a very private person" was item 8 and Form 3, on which this was item 1. Item 2–10–3 exhibits significant DIF for the second contrast, comparing Form 1 (on which "I keep in the background" was the second item) with Form 2, where nine items preceded it. Table 18.12 shows the

Table 18.10 Estimated Item Parameters and Their Standard Errors for "E" Item 1-1-6

Item	Form	a	s.e.	b_1	s.e.	b_2	s.e.	b_3	s.e.	b_4	s.e.
1	1	2.18	0.23	–3.43	0.42	–2.39	0.21	–0.89	0.09	0.70	0.10
1	2	2.00	0.20	–3.17	0.33	–2.05	0.17	–0.61	0.09	0.68	0.10
6	3	1.90	0.20	–3.33	0.39	–1.63	0.14	–0.49	0.08	0.95	0.12

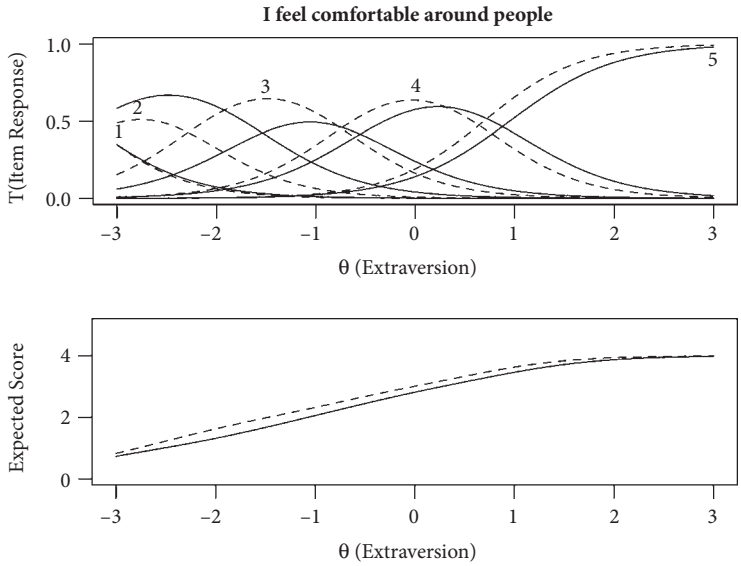


Figure 18.6 Trace lines (upper panel) and expected score curves (lower panel) for item 1-1-6; the dashed lines are for the average of the item parameters for Forms 1 and 2, while the solid lines are for the significantly different item parameters for Form 3.

parameter estimates for the three groups for these two items.

Figure 18.7 shows the differences between the fitted trace lines (upper panel) and expected score curves (lower panel) for item 8-8-1; the dashed lines are for the average of the item parameters for Forms 1 and 2, while the solid lines are for the significantly different item parameters for Form 3. For this item, the effect of "being first" is symmetrical: more extreme values of θ (Extraversion) are required to select either response 1 or 5, and to a lesser extent 2 or 4. The result for the expected scores as functions of θ (Extraversion), in the lower panel of Figure 18.7, gives the appearance of a difference in discrimination, although the difference in the parameters is entirely in the thresholds.

Figure 18.8 shows the differences between the fitted trace lines (upper panel) and expected score

Table 18.11 DIF Statistics for the Analysis of the Five "I" Items

Item numbers in:											
Group 1	Group 2	Group 3	Contrast	Total X^2	*df*	*P*	X^2_a	*df*	*p*	$X^2_{c\|a}$	*df*	*p*
8	8	1	1	32.4	5	.0001	1.9	1	.1724	30.6	4	.0001
			2	6.2	5	.2879	0.2	1	.6966	6.1	4	.1946
9	9	2	1	2.5	5	.7778	0.2	1	.6378	2.3	4	.6864
			2	8.6	5	.1267	1.2	1	.2790	7.4	4	.1157
2	10	3	1	0.9	5	.9729	0.0	1	.8666	0.8	4	.9338
			2	38.7	5	.0001	8.3	1	.0039	30.4	4	.0001
3	11	4	1	9.7	5	.0851	1.8	1	.1813	7.9	4	.0962
			2	15.4	5	.0087	5.9	1	.0149	9.5	4	.0496
6	12	5	1	11.7	5	.0386	1.5	1	.2248	10.3	4	.0363
			2	6.4	5	.2690	2.4	1	.1233	4.0	4	.4027

Table 18.12 Estimated Item Parameters and Their Standard Errors for "I" Items 8-8-1 and 2-10-3

Item 8-8-1											
Item	**Form**	**a**	**s.e.**	**b_1**	**s.e.**	**b_2**	**s.e.**	**b_3**	**s.e.**	**b_4**	**s.e.**
8	1	1.10	0.15	−1.88	0.24	−0.72	0.14	0.67	0.13	2.27	0.28
8	2	1.18	0.14	−1.65	0.20	−0.34	0.12	0.92	0.14	2.43	0.27
1	3	0.91	0.13	−2.78	0.39	−0.44	0.14	1.48	0.23	4.15	0.61
Item 2-10-3											
Item	**Form**	**a**	**s.e.**	**b_1**	**s.e.**	**b_2**	**s.e.**	**b_3**	**s.e.**	**b_4**	**s.e.**
2	1	1.65	0.20	−3.12	0.35	−1.31	0.14	−0.06	0.08	1.83	0.18
10	2	2.64	0.28	−2.18	0.19	−1.30	0.12	−0.13	0.08	0.81	0.09
3	3	2.10	0.25	−2.54	0.25	−1.41	0.13	−0.12	0.08	1.17	0.11

curves (lower panel) for item 2–10–3; the solid lines are for the item parameters for Form 1, while the dashed lines are for the significantly different item parameters for Form 2. As was the case for item 8–8-1, the effect of "being second" (here) is symmetrical: more extreme values of θ (Extraversion) are required to select either response 1 or 5, and to a lesser extent 2 or 4, compared to the same item presented tenth. Again, the result for the expected scores as functions of θ (Extraversion), in the lower panel of Figure 18.7, gives the appearance of a difference in discrimination; the statistical tests in Table 18.11 show that the significant effects are on both the discrimination and threshold parameters.

CONCLUSION FROM DIF ANALYSIS OF ITEM POSITIONING

For the most part, the items can be moved without affecting the responses, but some differences happen, especially for items in the first positions. This effect of serial position on item responses was initially explored by Knowles (1988), and Steinberg (1994) applied the methods of IRT to uncover a specific effect of context on item responses that is similar to the ones found in the analyses described

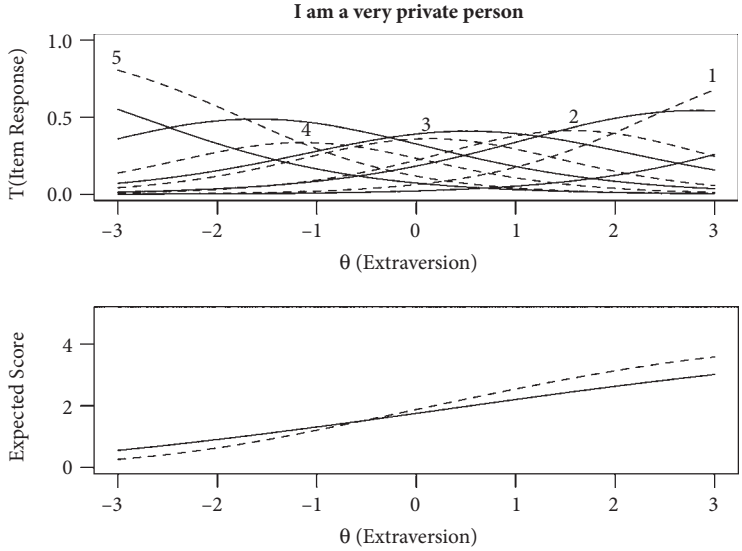


Figure 18.7 Trace lines (upper panel) and expected score curves (lower panel) for item 8-8-1; the dashed lines are for the average of the item parameters for Forms 1 and 2, while the solid lines are for the significantly different item parameters for Form 3.

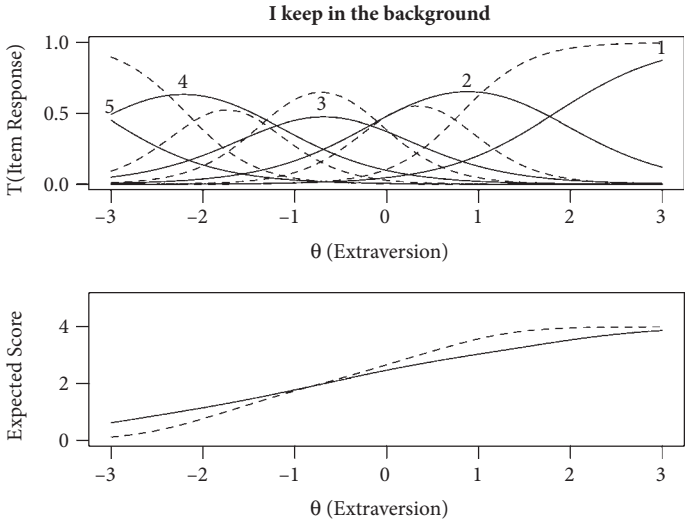


Figure 18.8 Trace lines (upper panel) and expected score curves (lower panel) for item 2-10-3; the solid lines are for the item parameters for Form 1; the dashed lines are for the significantly different item parameters for Form 2.

here. Answering prior questions may facilitate construct-related item interpretation, information relevant to the item may become more accessible, or an overarching construct-related judgment about the self may be developed. These processes may be responsible for the differences found between items presented early compared to later in the questionnaire.

Multidimensional Item Response Theory

The Multidimensional Logistic Model for Dichotomous Responses

Until relatively recently, applications of IRT were limited to measurement of unidimensional θ—that is, a single construct. This was largely a computational limitation; the conceptual generalization of IRT to *multidimensional item response theory* (MIRT) has been the subject of active research for some time. Reckase (2009) has provided a book-length treatment of the topic; the brief discussion that follows describes and illustrates only a few essential topics.

IRT becomes MIRT when more than one latent variable is required to account for the observed pattern of covariation among item responses. It is conventional to refer to the collection of several latent variables as $\boldsymbol{\theta}$; the bold typeface indicates that represents a vector with as many values as there are latent dimensions. In this chapter, we consider only *compensatory* MIRT models, which assume that the probability of an item response depends on a linear combination of the component values of $\boldsymbol{\theta}$. It is convenient to represent that linear combination using vector multiplication, in which $\mathbf{a}'\boldsymbol{\theta} = a_1\theta_1 + a_2\theta_2 + \cdots + a_p\theta_p$ for a p-dimensional latent variable. Using that notation, the trace surface for the multidimensional generalization of the 2PL model is

$$T_i(u_i = 1) = \frac{1}{1 + \exp[-(\mathbf{a}_i'\boldsymbol{\theta} + c_i)]}. \qquad (11)$$

In equation (11), T is the surface over the $\boldsymbol{\theta}$ (hyper) plane that traces the probability of a positive response ($u_i = 1$) to item i. For two-dimensional models, $\boldsymbol{\theta}$ is a two-dimensional plane with values of θ_1 from low to high in one direction, and values of θ_2 from low to high in the other (orthogonal) direction; T is a surface that rises from zero (for low values of θ_1 and θ_2) to one (for high values of θ_1 and θ_2 or both). Models with more than two latent dimensions are difficult to visualize; however, considering the components one at a time, such models can nonetheless be useful.

Comparing equation (11) with equation (1), the only difference is that the notation $\mathbf{a}'$ represents a vector of as many slope or discrimination parameters as there are dimensions, and $\boldsymbol{\theta}$ is a vector of the same dimensionality of latent variable values. So relative to equation (1), equation (11) adds one more parameter (one more a value) to the model for each additional dimension. The a parameters each measure the degree of relation of the item response with the corresponding dimension of $\boldsymbol{\theta}$ (that is, with each construct). The scalar-valued parameter c remains an intercept parameter that reflects the

overall probability of endorsement of the item. In contrast to the unidimensional 2PL model, there is no alternate formulation of the model with a threshold parameter, because for a MIRT model the threshold (the location of $T = 0.5$) is a line for a two-dimensional model, or a (hyper)plane for higher-dimensional models—there is no scalar-valued threshold.

Multidimensional Logistic Models for Polytomous Responses

The graded model is constructed of 2PL components, so substitution of trace surface models of the form of equation (11) into equation (9) yields a multidimensional graded logistic model:

$$T_i(u_i = k) = \frac{1}{1 + \exp[-\mathbf{a}_i'\theta + c_i]} - \frac{1}{1 + \exp[-\mathbf{a}_i'\theta + c_{i+1}]}. \quad (12)$$

The model in equation (12) is like that in equation (9), except that there are as many slope (a) parameters as there are dimensions. In a graded model, there are $K - 1$ intercept parameters for an item with K response categories.

MIRT Is Item Factor Analysis

Compensatory MIRT models are factor analysis models for categorical item response data, or *item factor analysis* models (Bartholomew & Knott, 1999; Bock, Gibbons, & Muraki, 1988; Bock & Moustaki, 2007; Wirth & Edwards, 2007). The differences between MIRT and factor analysis involve the use of nonlinear models for the probability of a categorical item response in MIRT versus linear models for continuous scores in factor analysis, and the fact that the regression parameters measuring the association between the observed variables and the latent variables are reported as a parameters in MIRT and (conventionally) *factor loadings*, λ, in factor analysis. The former is unavoidable given the nature of the observed data. The latter is a matter of convention; either MIRT slope parameters or factor loadings can be reported using the relations

$$\lambda = \frac{a/1.7}{\sqrt{1 + \sum\left(\frac{a}{1.7}\right)^2}} \quad (13)$$

and

$$a = 1.7\frac{\lambda}{\sqrt{1 - \sum \lambda^2}} \quad (14)$$

for orthogonal factors (Bock, Gibbons, & Muraki, 1988, pp. 263–264; McLeod, Swygert, & Thissen, 2001, p. 199).

The Relation of the Model with the Data, and Parameter Estimation

When multidimensional trace surfaces, and a multidimensional population distribution, are substituted into equation (3), estimation of the item parameters proceeds just as it does for unidimensional models, after some set of constraints is imposed on the model for identification.

When the trace surface models take the form of equations (11) or (12), with more than one latent variable θ and correspondingly more than one a parameter per item, the parameters of the model are not identified due to what factor analysts call rotational indeterminacy (Harmon, 1967, p. 23). An infinite number of collections of **a** vectors yield the same fit to the data; the differences among those **a** vectors correspond to different orientations of the reference axes $\boldsymbol{\theta}$. This indeterminacy can be resolved in one of two ways to compute a unique set of item parameters: An informatively restricted model can be used and the results interpreted directly, or a minimally restricted model can be used and the results *rotated* into interpretable form.

Model Identification: Restriction or Rotation

RESTRICTED, OR CONFIRMATORY FACTOR ANALYSIS, MODELS

Jöreskog (1966) proposed that the problem of rotational indeterminacy could be avoided, rather than solved, by fitting a multidimensional model with sufficient restrictions on the parameters to identify the model. At first this procedure was called *restricted* factor analysis (Jöreskog & Gruvaeus, 1967), but it was soon renamed *confirmatory factor analysis* (CFA) (Jöreskog, 1969) to emphasize its potential use for hypothesis testing. The nomenclature CFA is now in near-universal usage, so we refer to restricted models as "CFA models," whether or not the context is hypothesis testing.

"Correlated Simple Structure" or "Independent Clusters" Models

One class of CFA models that is widely useful in item factor analysis specifies that the test measures more than one latent variable, but that each item serves as an indictor for only one of the constructs; further, the constructs may be correlated. This type of model involves a pattern of slopes

(or equivalently, factor loadings) that are estimated or fixed at the value zero in a pattern that looks like

x	0	0
x	0	0
x	0	0
0	x	0
0	x	0
0	x	0
0	0	x
0	0	x
0	0	x

for nine items that measure three constructs, where x represents an estimated slope and 0 means the corresponding $a = 0.0$. For example, one could have an emotional distress scale with items measuring depressive symptoms, anxiety, and anger. The data analyst knows in advance which items are intended to measure which of the three constructs and would restrict the *a* parameters for each item on the other two θs to be zero. Such restrictions overidentify the model, even when the correlations among the three constructs are also non-zero estimated values. Such models are variously called *correlated simple structure* or *independent clusters* or *perfect clusters* models; they express an extreme form of Thurstone's (1947) *simple structure*. With these models, scores that are estimates of the level on the θs are straightforwardly interpreted—in the example, they are scores for depressive symptoms, anxiety, and anger.

More generally, CFA models may be somewhat less restricted; for example, an item may have more than one estimated (non-zero) slope parameter, if it measures more than one of the constructs. For *m* factors, as long as there are at least $m - 1$ zero loadings per column, the model is identified when the factor intercorrelations are estimated (Browne, 2001); if the interfactor correlations are all constrained to be zero, the model is *orthogonal* and only $m(m - 1)/2$ suitably placed zeros are required for identification.

Bifactor Models

Another class of CFA models that is useful for item factor analysis includes *bifactor* (Holzinger & Swinford, 1937) or *hierarchical* models. This type of model involves a pattern of slopes (or equivalently, factor loadings) that are estimated or fixed at the value zero in a pattern that looks like

X	x	0	0
X	x	0	0
X	x	0	0
X	0	x	0
X	0	x	0
X	0	x	0
X	0	0	x
X	0	0	x
X	0	0	x

again for nine items that measure three constructs, where x represents an estimated slope and 0 means the corresponding $a = 0.0$. In bifactor models the factor intercorrelations are restricted to be zero (the θs are orthogonal). To use the same example, one could have an emotional distress scale with items measuring depressive symptoms, anxiety, and anger. In this parameterization, the first factor (θ, or construct) is generalized emotional distress. The second θ is a different kind of construct: it is an individual differences variable that measures the degree to which the respondent is relatively higher or lower specifically on items measuring depressive symptoms, given his or her level on global emotional distress. The third and fourth factors similarly represent deviations for anxiety and anger. Scores derived from bifactor models have different interpretations depending on whether they are for the first general factor or for the second-tier factors. The score on the first, general factor would straightforwardly be a level of emotional distress in this example. However, the scores on the cluster-specific factors are not "depressive symptoms, anxiety, and anger" scores; they are *residual* or *deviation* scores for higher or lower levels of those constructs over and above the general factor score, which already includes a concatenation of all three constructs. If one wanted to derive three scores with the labels "depressive symptoms, anxiety, and anger" from a bifactor model, one would have to compute three linear combinations of the general factor scores with each of the three second-tier scores.

The bifactor model has been widely used in item factor analysis since Gibbons and Hedeker (1992) provided a practical, efficient estimation algorithm for dichotomous models; Gibbons, Bock, Hedeker, Weiss, Segawa, Bhaumik, and colleagues (2007) extended this algorithm to polytomous models.

Cai, Yang, and Hansen (2011) have recently provided a number of illustrations of the usefulness of the bifactor model in item factor analysis, and Cai (2010) has extended Gibbons and Hedeker's computational results to cases in which the "general" part of the model contains more than one factor.

Bifactor models provide a computationally efficient way to test the hypothesis that clusters of items on a scale exhibit LD, which means that they measure their own cluster factor to some degree, in addition to a more general construct, and/or measure the relative degree to which items measure cluster constructs versus a general construct.

The Testlet Response Model

Wainer, Bradlow, and Wang (2007) summarize a decade of research with the *testlet response model* that measures the effects of clustering or secondary factors in measurement instruments. In the Wainer-Bradlow-Wang parameterization, that model looks somewhat different than a CFA model. Each item has only one slope parameter, and variance parameters are estimated—the variances of individual differences on the second-tier factors, relative to 1.0 as the (arbitrarily defined) variance of the general factor. However, Li, Bolt, and Fu (2006) showed that the testlet response model is a constrained bifactor model (see also Thissen & Steinberg, 2010). If the slope parameters on the second-tier factors are constrained to be equal for each item to that item's general-factor slope, and the variance of the factor is estimated instead of fixed at a reference value of 1.0, the bifactor model becomes the testlet response model.

Further, the testlet response model, the bifactor model, and the correlated independent clusters model are much more closely related than their disparate descriptions might suggest. For traditional factor analysis, extending results obtained by Schmid and Leiman (1957), Yung, Thissen, and McLeod (1999) showed that a second-order factor model is equivalent to a constrained bifactor model; the constraints are the same as those that give rise to the testlet response model. Rijmen (2010) has shown that the equivalence of the testlet response model, a constrained bifactor model, and the second-order factor model also applies to MIRT models for categorical item response data. A correlated simple structure model becomes a second-order factor model when a smaller one-factor model is nested within the main model, to explain the correlations among the factors. The upshot of all of this is that if a second-order factor is sufficient to explain the interfactor correlations of a correlated simple structure model, that model is hierarchically nested within a bifactor model, because it is a constrained version of the bifactor model. This set of relations explains the common alternative use of simple structure or bifactor MIRT models with the same, or similar, data. For a much more extensive discussion of the use of bifactor and related models in personality measurement, see Reise, Moore, and Haviland (2010).

An Example: The PROMIS Pediatric Emotional Distress Measures

THE DATA

The data for this example used responses to the PROMIS Pediatric Emotional Distress item banks for Depressive Symptoms and Anxiety (Irwin, Stucky, Thissen, DeWitt, Lai, Yeatts, Varni, & DeWalt, 2010) and Anger (Irwin, Stucky, Langer, Thissen, DeWitt, Lai, Yeatts, Varni, & DeWalt, 2012). The respondents were 1,425 children and adolescents, ages 8 to 17, recruited from treatment centers for six chronic conditions (cancer, chronic kidney disease, obesity, rehabilitative needs, rheumatic disease, sickle cell disease). Some randomly assigned subsets of the sample were administered the entire PROMIS Pediatric Emotional Distress item banks (14 Depressive Symptoms items, 15 Anxiety items, and 6 Anger items); other subsets were administered short forms of the scales, 8 items each for Depressive Symptoms and Anxiety, with the 6 Anger items. The data were collected as part of a multipurpose validity study. All items used five response alternatives (0 = never, 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always), and were fitted with unidimensional and multidimensional versions of the graded item response model.

INDEPENDENT CLUSTERS CFA

We first use these data to illustrate correlated simple structure or independent clusters confirmatory item factor analysis. The primary purposes of this analysis are (a) to test the hypothesis that a correlated three-factor model (one factor each for Depressive Symptoms, Anxiety, and Anger) fits the data better than a unidimensional "emotional distress" model and (b) to estimate the correlations among the three hypothesized latent variables.

Table 18.13 lists item identification codes for the 35 items, the item stems, and the slope (a) parameters and their standard errors for the confirmatory three-factor solution. The parameter values

Table 18.13 Slope Parameters and Their Standard Errors for the Three-Factor CFA Model for the PROMIS Pediatric Emotional Distress Data

Item ID	Item Stem	a_1	s.e.	a_2	s.e.	a_3	s.e.
Ang1–1	I felt mad.	2.84	0.30	0.0	—	0.0	—
Ang1–5	I was so angry I felt like yelling at somebody.	2.67	0.28	0.0	—	0.0	—
Ang1–10	I felt upset.	2.50	0.24	0.0	—	0.0	—
Ang1–3	I was so angry I felt like throwing something.	2.46	0.28	0.0	—	0.0	—
Ang1–9	I felt fed up.	2.20	0.21	0.0	—	0.0	—
Ang1–8	When I got mad, I stayed mad.	2.13	0.26	0.0	—	0.0	—
Anx2–2	I felt scared.	0.0	—	3.15	0.21	0.0	—
Anx2–5	I worried when I was at home.	0.0	—	2.97	0.36	0.0	—
Anx2–9	I felt worried.	0.0	—	2.93	0.18	0.0	—
Anx2–1	I felt like something awful might happen.	0.0	—	2.87	0.18	0.0	—
Anx2–4	I worried when I went to bed at night.	0.0	—	2.60	0.18	0.0	—
Anx1–3	I worried about what could happen to me.	0.0	—	2.46	0.15	0.0	—
Anx1–8	I felt nervous.	0.0	—	2.37	0.15	0.0	—
Anx2–6	I thought about scary things.	0.0	—	2.16	0.14	0.0	—
Anx2–3	I was worried I might die.	0.0	—	2.13	0.25	0.0	—
Anx1–1	I got scared really easy.	0.0	—	2.12	0.21	0.0	—
Anx1–5	I woke up at night scared.	0.0	—	2.03	0.22	0.0	—
Anx1–7	I was afraid that I would make mistakes.	0.0	—	1.88	0.13	0.0	—
Anx1–6	I worried when I was away from home.	0.0	—	1.86	0.19	0.0	—
Anx2–7	I was afraid of going to school.	0.0	—	1.87	0.23	0.0	—
Anx1–9	It was hard for me to relax.	0.0	—	1.52	0.15	0.0	—
Dep1–7	I felt everything in my life went wrong.	0.0	—	0.0	—	3.05	0.22
Dep2–3	I felt sad.	0.0	—	0.0	—	2.91	0.19
Dep2–5	I thought that my life was bad.	0.0	—	0.0	—	2.90	0.19
Dep1–4	I felt alone.	0.0	—	0.0	—	2.89	0.20
Dep2–10	I felt lonely.	0.0	—	0.0	—	2.79	0.19
Dep2–11	I felt unhappy.	0.0	—	0.0	—	2.62	0.15
Dep2–7	I could not stop feeling sad.	0.0	—	0.0	—	2.57	0.18
Dep1–5	I felt like I couldn't do anything right.	0.0	—	0.0	—	2.40	0.15

(*continued*)

Table 18.13 (Continued)

Item ID	Item Stem	a_1	s.e.	a_2	s.e.	a_3	s.e.
Dep1–8	Being sad made it hard for me to do things with my friends.	0.0	—	0.0	—	2.10	0.22
Dep2–1	I felt too sad to eat.	0.0	—	0.0	—	1.87	0.24
Dep2–6	It was hard for me to have fun.	0.0	—	0.0	—	1.74	0.18
Dep2–8	I felt stressed.	0.0	—	0.0	—	1.68	0.16
Dep2–2	I didn't care about anything.	0.0	—	0.0	—	1.41	0.16
Dep1–1	I wanted to be by myself.	0.0	—	0.0	—	0.92	0.12

The items are sorted within factor by the values of *a*.

listed as 0.0 with no standard errors in Table 18.13 are fixed values. Intercept parameters were also estimated for each item; however, those are not involved in the interpretation of the primary results for this example, so they are not tabulated here. This three-dimensional model fits the data significantly better than a one-dimensional model that has only a single "emotional distress" latent variable: −2loglikelihood for the three-dimensional model is 50666.56, for the unidimensional model it is 52328.28, and the difference, 1661.72, is distributed as χ^2 on 3 *df* under the unidimensional null hypothesis. That is highly significant, so we reject unidimensionality in favor of the three-dimensional model.

Table 18.14 has the same structure as Table 18.13, except that the IRT slope (*a*) parameters have been converted to standardized factor loadings (λ) to aid their interpretation for readers more familiar with factor analytic models expressed in that way. Table 18.15 shows the estimates of the correlations among the three latent variables (the Depressive Symptoms, Anxiety, and Anger constructs). These values have the same estimand as "disattenuated" estimates of correlation using summed score theory—they are estimates of the correlations among the latent variables, not among observed scores. The estimates of the correlations between both Anger and Anxiety and Depressive Symptoms are 0.78, while the correlation between Anger and Anxiety is somewhat lower, 0.66. This is an example of results that may be obtained measuring three highly correlated but nevertheless distinct constructs.

BIFACTOR ANALYSIS

These data can also be used to illustrate the use of a bifactor (or hierarchical) model to investigate multidimensionality. The left block of columns of Table 18.16 lists the slope parameters for a general ("emotional distress") factor (a_1) and three second-tier factors (for the Anger, Anxiety, and Depressive Symptoms items; a_2, a_3, and a_4 respectively). The fact that the cluster-specific slope values for the Anger and Anxiety items (a_2 and a_3) are all substantially larger than their standard errors leads to the same conclusion as the three-factor simple-structure model: The item set is multidimensional, with the Anger, Anxiety, and Depressive Symptoms items measuring somewhat distinct constructs. In this analysis, it is a curiosity, but not unusual, that most of the second-tier slopes for the Depressive Symptoms items (a_3) are not larger than twice their standard errors. There is evidence of a locally dependent doublet (items Dep1–4, "I felt alone" and Dep2–10, "I felt lonely"); one aspect of what has happened in this analysis is that the doublet has engaged in "θ theft" (Thissen & Steinberg, 2010), meaning that the (trivial) construct that explains excessive dependence between those two items has stolen θ_4. The second thing that is probably happening is that, given this particular set of items, the general factor *is* the Depressive Symptoms construct. The Anger and Anxiety second-tier constructs explain some additional covariation among their clusters of items, but the general factor explains all of the covariation among the Depressive Symptoms items (except for the doublet). This result should not be overinterpreted substantively; other analyses of other subsets of these items have the Anxiety cluster taking over the general factor (Irwin, Stucky, Thissen, DeWitt, Lai, Yeatts, Varni, & DeWalt, 2010). The relative numbers of items in the clusters, and their discrimination, work together to determine the orientation of the general factor. Nevertheless, the overall conclusion is that the item set is three-dimensional, not unidimensional.

Table 18.14 Factor Loadings and Their Standard Errors for the Three-Factor CFA Model for the PROMIS Pediatric Emotional Distress Data

Item ID	Item Stem	λ_1	s.e.	λ_2	s.e.	λ_3	s.e.
Ang1–1	I felt mad.	0.86	0.04	0.00	—	0.00	—
Ang1–5	I was so angry I felt like yelling at somebody.	0.84	0.04	0.00	—	0.00	—
Ang1–10	I felt upset.	0.83	0.04	0.00	—	0.00	—
Ang1–3	I was so angry I felt like throwing something.	0.82	0.05	0.00	—	0.00	—
Ang1–9	I felt fed up.	0.79	0.05	0.00	—	0.00	—
Ang1–8	When I got mad, I stayed mad.	0.78	0.06	0.00	—	0.00	—
Anx2–2	I felt scared.	0.00	—	0.88	0.02	0.00	—
Anx2–5	I worried when I was at home.	0.00	—	0.87	0.04	0.00	—
Anx2–9	I felt worried.	0.00	—	0.87	0.02	0.00	—
Anx2–1	I felt like something awful might happen.	0.00	—	0.86	0.02	0.00	—
Anx2–4	I worried when I went to bed at night.	0.00	—	0.84	0.03	0.00	—
Anx1–3	I worried about what could happen to me.	0.00	—	0.82	0.03	0.00	—
Anx1–8	I felt nervous.	0.00	—	0.82	0.03	0.00	—
Anx2–6	I thought about scary things.	0.00	—	0.79	0.03	0.00	—
Anx2–3	I was worried I might die.	0.00	—	0.78	0.06	0.00	—
Anx1–1	I got scared really easy.	0.00	—	0.78	0.05	0.00	—
Anx1–5	I woke up at night scared.	0.00	—	0.77	0.06	0.00	—
Anx1–7	I was afraid that I would make mistakes.	0.00	—	0.74	0.04	0.00	—
Anx1–6	I worried when I was away from home.	0.00	—	0.74	0.06	0.00	—
Anx2–7	I was afraid of going to school.	0.00	—	0.74	0.07	0.00	—
Anx1–9	It was hard for me to relax.	0.00	—	0.67	0.06	0.00	—
Dep1–7	I felt everything in my life went wrong.	0.00	—	0.00	—	0.87	0.03
Dep2–3	I felt sad.	0.00	—	0.00	—	0.86	0.02
Dep2–5	I thought that my life was bad.	0.00	—	0.00	—	0.86	0.02
Dep1–4	I felt alone.	0.00	—	0.00	—	0.86	0.03
Dep2–10	I felt lonely.	0.00	—	0.00	—	0.85	0.03
Dep2–11	I felt unhappy.	0.00	—	0.00	—	0.84	0.02
Dep2–7	I could not stop feeling sad.	0.00	—	0.00	—	0.84	0.03
Dep1–5	I felt like I couldn't do anything right.	0.00	—	0.00	—	0.82	0.03
Dep1–8	Being sad made it hard for me to do things with my friends.	0.00	—	0.00	—	0.78	0.06
Dep2–1	I felt too sad to eat.	0.00	—	0.00	—	0.74	0.07

(*continued*)

Table 18.14 (Continued)

Item ID	Item Stem	λ_1	s.e.	λ_2	s.e.	λ_3	s.e.
Dep2–6	It was hard for me to have fun.	0.00	—	0.00	—	0.72	0.06
Dep2–8	I felt stressed.	0.00	—	0.00	—	0.70	0.06
Dep2–2	I didn't care about anything.	0.00	—	0.00	—	0.64	0.07
Dep1–1	I wanted to be by myself.	0.00	—	0.00	—	0.48	0.08

THE TESTLET RESPONSE MODEL

Yet another way to analyze these data is to use the testlet response model (Wainer, Bradlow, & Wang, 2007). For direct comparison with the results from the bifactor and independent clusters analysis, that model has been fitted to these data using ML estimation with the software IRTPRO (Cai, Thissen, &du Toit,, 2011), with the results shown in the right block of Table 18.16. Wang, Bradlow, and Wainer's (2005) SCORIGHT software uses a Bayesian MCMC algorithm to fit this same model—that produces more information but is more complex to use.

The testlet response model is like the bifactor model, except that the slopes on the second-tier factors are set equal to the slopes on the general factor, as shown in the table. Then the variances of the second-tier factors are estimated, relative to 1.0, which is the scale-defining fixed variance of the general factor. The variance estimates for θ_2 and θ_3 are 0.51 and 0.54, respectively, indicating that there is substantial individual differences variation remaining for the Anger and Anxiety constructs after covariation among the item responses due to global emotional distress has been accounted for. The variance estimate for θ_4 (for the Depressive Symptoms cluster) is only 0.08, again indicating that in this analysis the general factor (θ_1) *is* the Depressive Symptoms construct.

Using these same data we have shown three analyses that all lead to the same conclusion: the item set is three-dimensional and is likely best divided into three scales to yield three scores for Anger, Anxiety, and Depressive Symptoms. In any particular context, one or another of the models we have used may be more numerically stable in estimation, and/or may give clearer results. For a much more thorough investigation of the general and specific components of depression and anxiety for adults using the bifactor model, see Simms, Grös, Watson, and O'Hara (2008); for another comparison of the same varieties of CFA model fitting used in this example, see Reise, Morizot, and Hays (2007).

Table 18.15 Correlations Among the Factors and Their Standard Errors for the Three-Factor CFA Model for the PROMIS Pediatric Emotional Distress Data

	θ_1	s.e.	θ_2	s.e.
Anger – θ_1	1.00			
Anxiety – θ_2	0.66	0.03	1.00	
Depressive symptoms – θ_3	0.78	0.02	0.78	0.02

An Example: The 6-Item BIS

THE DATA

The item response data were obtained from the same respondents and BIS impulsiveness measure described previously, in the section introducing the graded model. In addition to the five items used in the earlier example, this analysis adds one more item (BIS12, "I am a careful thinker") to the analysis to form a six-item set.

BIS 6-ITEM ANALYSIS

Unidimensional graded model analysis of the six-item set reveals a substantial standardized *LD* X^2 index (value = 19.8) between the items BIS8 "I am self controlled" and BIS12 "I am a careful thinker." The presence of LD implies that there is more covariation between two items than is accounted for by the unidimensional IRT model—the item pair exhibits excess covariation; that indicates a violation of the assumption of local independence. Another perspective on LD is that it reflects another factor, but the excess covariation is considered a nuisance (or even an artifact of similarity or wording or meaning of the items) rather than a substantive factor on which one measures individual differences.

This six-item set will be used to illustrate methods to investigate and model LD between pairs or triplets of items. First, an analysis may be done to evaluate the significance of the LD. We add parameters

Table 18.16 Slope Parameters and Their Standard Errors for the Bifactor and Testlet Response Models for the PROMIS Pediatric Emotional Distress Data

	Bifactor Model								Testlet Response Model				
Item ID	a_1	s.e.	a_2	s.e.	a_3	s.e.	a_4	s.e.	a_1	s.e.	a_2	a_3	a_4
Ang1–1	2.11	0.26	1.68	0.24	0.0	—	0.0	—	2.31	0.22	2.31	0.0	0.0
Ang1–5	2.23	0.29	2.14	0.30	0.0	—	0.0	—	2.18	0.21	2.18	0.0	0.0
Ang1–10	2.22	0.24	1.01	0.16	0.0	—	0.0	—	2.03	0.21	2.03	0.0	0.0
Ang1–3	2.01	0.27	2.00	0.30	0.0	—	0.0	—	2.00	0.21	2.00	0.0	0.0
Ang1–9	1.79	0.19	1.15	0.18	0.0	—	0.0	—	1.79	0.18	1.79	0.0	0.0
Ang1–8	1.69	0.22	1.27	0.21	0.0	—	0.0	—	1.73	0.18	1.73	0.0	0.0
Anx2–2	2.45	0.19	0.0	—	2.30	0.20	0.0	—	2.56	0.20	0.0	2.56	0.0
Anx2–5	2.39	0.31	0.0	—	1.73	0.27	0.0	—	2.42	0.24	0.0	2.42	0.0
Anx2–9	2.26	0.16	0.0	—	1.92	0.16	0.0	—	2.38	0.19	0.0	2.38	0.0
Anx2–1	2.26	0.17	0.0	—	1.73	0.14	0.0	—	2.33	0.18	0.0	2.33	0.0
Anx2–4	2.02	0.16	0.0	—	1.66	0.15	0.0	—	2.11	0.17	0.0	2.11	0.0
Anx1–3	1.94	0.14	0.0	—	1.46	0.13	0.0	—	1.99	0.15	0.0	1.99	0.0
Anx1–8	1.80	0.14	0.0	—	1.62	0.14	0.0	—	1.93	0.15	0.0	1.93	0.0
Anx2–6	1.62	0.13	0.0	—	1.56	0.14	0.0	—	1.75	0.14	0.0	1.75	0.0
Anx2–3	1.81	0.22	0.0	—	1.07	0.19	0.0	—	1.73	0.18	0.0	1.73	0.0
Anx1–1	1.55	0.20	0.0	—	1.70	0.21	0.0	—	1.72	0.17	0.0	1.72	0.0
Anx1–5	1.61	0.21	0.0	—	1.30	0.20	0.0	—	1.65	0.17	0.0	1.65	0.0
Anx1–7	1.68	0.12	0.0	—	0.89	0.10	0.0	—	1.53	0.12	0.0	1.53	0.0
Anx1–6	1.39	0.18	0.0	—	1.28	0.19	0.0	—	1.51	0.15	0.0	1.51	0.0
Anx2–7	1.58	0.22	0.0	—	0.99	0.20	0.0	—	1.52	0.18	0.0	1.52	0.0
Anx1–9	1.40	0.15	0.0	—	0.68	0.14	0.0	—	1.24	0.13	0.0	1.24	0.0
Dep1–7	3.37	0.25	0.0	—	0.0	—	−0.35	0.18	2.93	0.18	0.0	0.0	2.93
Dep2–3	2.87	0.19	0.0	—	0.0	—	0.10	0.16	2.79	0.16	0.0	0.0	2.79
Dep2–5	3.10	0.21	0.0	—	0.0	—	−0.22	0.17	2.79	0.17	0.0	0.0	2.79
Dep1–4	4.49	0.62	0.0	—	0.0	—	2.63	0.55	2.77	0.17	0.0	0.0	2.77
Dep2–10	3.96	0.42	0.0	—	0.0	—	2.23	0.41	2.68	0.16	0.0	0.0	2.68
Dep2–11	2.58	0.15	0.0	—	0.0	—	0.16	0.14	2.52	0.14	0.0	0.0	2.52
Dep2–7	2.58	0.19	0.0	—	0.0	—	−0.05	0.16	2.47	0.15	0.0	0.0	2.47
Dep1–5	2.44	0.16	0.0	—	0.0	—	0.01	0.15	2.31	0.14	0.0	0.0	2.31
Dep1–8	2.04	0.23	0.0	—	0.0	—	0.24	0.19	2.01	0.17	0.0	0.0	2.01

(*continued*)

Table 18.16 (Continued)

	Bifactor Model								Testlet Response Model				
Item ID	a_1	s.e.	a_2	s.e.	a_3	s.e.	a_4	s.e.	a_1	s.e.	a_2	a_3	a_4
Dep2–1	1.83	0.25	0.0	—	0.0	—	0.31	0.22	1.80	0.18	0.0	0.0	1.80
Dep2–6	1.77	0.19	0.0	—	0.0	—	0.07	0.17	1.67	0.15	0.0	0.0	1.67
Dep2–8	1.71	0.17	0.0	—	0.0	—	−0.09	0.17	1.62	0.13	0.0	0.0	1.62
Dep2–2	1.45	0.17	0.0	—	0.0	—	−0.12	0.18	1.36	0.13	0.0	0.0	1.36
Dep1–1	0.94	0.12	0.0	—	0.0	—	−0.14	0.16	0.88	0.09	0.0	0.0	0.88
								Variance of θ:		0.51	0.54	0.08	

The items are sorted as in Tables 18.13 and 18.14.

to the model that account for the excess covariation and evaluate the significance of the additional parameters. To accomplish this, a bifactor model is fitted that estimates an equal-slope second factor composed of the pair of items that show LD; this two-factor model is analogous to correlating the unique error variances between two items in confirmatory factor analysis. Table 18.17 lists the item parameter estimates obtained from this analysis. To evaluate the significance of the LD, one may calculate the likelihood ratio goodness-of-fit difference test, subtracting −2loglikelihood obtained from the bifactor analysis from the −2loglikelihood obtained from the six-item analysis. The result is $G^2(1) = 89.84$, $p < .0001$, indicating significant LD between these two items. Alternatively, one can compute the ratio of the bifactor slope estimate to its standard error, 1.38/0.16 = 8.62; for large samples that is distributed as a *z*-statistic, so $p < .0001$.

There can be different responses to evidence of significant LD depending on the goals of the analysis, several of which are simpler than fitting the testlet response model as was done in the previous example. If there is little concern about reducing the number of items on the instrument, then one solution is to eliminate one of the items of the LD pair. Selecting which item to retain may be made on the basis of considerations of item content or the magnitude of the slope. If one of the items has a substantially higher slope estimate than the other, and examination of the item content supports the idea that the item is a better indicator, the analyst might set aside the lower-slope item. Alternatively, without inspecting the item parameters, one may decide that one of the items seems to be more central to the intended construct or is worded in a more desirable way. For example, "I am self controlled" seems more central to the construct of impulsiveness than "I am a careful thinker."

If the goal is to retain all of the items, one may create a testlet of the LD pair (Steinberg & Thissen, 1996; Thissen & Steinberg, 2010). In this example, a testlet is a "super item" composed of the sum of the item responses for the two items. The sum of the two four-category (0, 1, 2, 3) items can be recoded to become a single item with seven categories (0–6). Table 18.18 lists the item parameters for the graded model fitted to the testlet created by summing

Table 18.17 Bifactor Model Parameter Estimates for Six BIS Items

Item	a_1	s.e.	a_2	s.e.	c_1	s.e.	c_2	s.e.	c_3	s.e.
BIS2	2.41	0.21	0.00	—	1.34	0.14	−3.21	0.22	−6.33	0.41
BIS5	1.03	0.09	0.00	—	0.92	0.08	−1.83	0.10	−4.02	0.19
BIS8	1.15	0.10	1.38	0.16	0.61	0.09	−2.45	0.14	−5.29	0.29
BIS12	1.36	0.12	1.38	0.16	1.44	0.11	−1.81	0.13	−5.55	0.34
BIS14	1.53	0.11	0.00	—	1.44	0.10	−1.89	0.11	−4.04	0.20
BIS19	1.37	0.10	0.00	—	2.13	0.11	−1.32	0.09	−3.63	0.17

Table 18.18 Graded Model Parameter Estimates for Six BIS Items, with BIS8 and BIS12 Responses Summed to Become One Testlet

Item	a	s.e.	b_1	s.e.	b_2	s.e.	b_3	s.e.	b_4	s.e.	b_5	s.e.	b_6	s.e.
BIS2	2.43	0.22	−0.55	0.05	1.33	0.07	2.61	0.15	—	—	—	—	—	—
BIS5	1.04	0.09	−0.89	0.09	1.77	0.14	3.88	0.31	—	—	—	—	—	—
BIS14	1.52	0.11	−0.95	0.07	1.24	0.08	2.65	0.16	—	—	—	—	—	—
BIS19	1.36	0.10	−1.56	0.10	0.97	0.08	2.66	0.17	—	—	—	—	—	—
BIS8plus12	1.12	0.09	−1.63	0.12	−0.21	0.06	1.20	0.10	2.21	0.15	4.01	0.31	5.00	0.44

the responses to items BIS8 and BIS12 and the other four items. The model exhibits adequate fit ($M_2(121) = 170.76$, $p = .002$; RMSEA = 0.02), and there are no large values of the *LD* X^2 indices.

In practice, researchers typically sum the responses to all of the items to obtain a score on the measure. Because the testlet represents the sum of the two items, summing all the items as usual gives the same result as summing four items and the testlet together. Thus, the testlet has the advantage of accounting for the LD without the loss of the item, and without altering interpretation of the traditionally used summed score.

EXPLORATORY ITEM FACTOR ANALYSIS

Before restricted or confirmatory factor analysis was developed in the 1960s, factor analysts used a two-stage procedure to obtain an identified factor analytic model with interpretable results. In the first stage, estimates of factor loadings were computed using an arbitrary minimal set of restrictions, and then in a second stage the loadings were *rotated* to become interpretable. Before the 1960s, this two-stage procedure was called simply *factor analysis*; after the development of CFA, it has become conventional to refer to the two-stage procedures as *exploratory factor analysis* (EFA).

The principal advantage of EFA is that it does not require *a priori* specification of the structure of the MIRT model. The principal disadvantages of EFA in the item factor analysis context are that it may require an excessive number of factors to fit the item response data, and rotation is required, after which the data analyst must remember that there are an infinite number of other, different, rotated solutions that fit the data equally well. EFA at the item level may require an excessive number of factors to fit because item response data may include several pairs or triples of items that are locally dependent—*doublets* or *triplets*. Expressed as an MIRT model, each doublet or triplet adds another factor; in a CFA framework, that is not an insurmountable problem, because the rest of the slopes/loadings on each of those added factors are zero, so few parameters are added. However, in EFA each additional factor adds almost as many slopes/loadings to the model as there are items. It is a general principle of statistical estimation that when more parameters are estimated, all parameters are estimated more poorly. In many cases, these phenomena render item-level EFA challenging or useless. However, in some cases it can be helpful to find unanticipated structure.

The first stage of EFA is usually done with orthogonal (uncorrelated) reference axes (θs). In an orthogonal model, minimal restriction for identification for m factors requires that $m(m-1)/2$ slopes (or loadings) be fixed at some specified value(s). Often this is done by fixing the slope(s) for the first item on the second factor, the first two items on the third factor, the first three items on the fourth factor, and so forth. These values may be fixed at zero, or, alternatively, to values obtained with some preliminary factor analysis of the interitem correlations; the software IRTPRO (Cai, Thissen, & du Toit, 2011) does the latter. ML estimates of the MIRT model parameters are then obtained using an arbitrary (uninterpretable) rotation of the reference axes.

After the ML estimates of the item parameters have been computed, the axes are rotated into an interpretable configuration. Historically, a good deal of research has developed effective analytic methods to rotate (or *transform*) factor loadings into regression coefficients on an interpretable set of reference axes, which are interpreted as the latent variables (Browne, 2001). An effective way to obtain a solution that most closely approximates a correlated simple structure or independent clusters model is oblique CF-Quartimax (Browne, 2001; Crawford & Ferguson, 1970).

In MIRT EFA, one does the ML estimation of the parameters as slopes (a) and intercepts (c),

because that is the parameterization for which IRT estimation is best understood and most commonly implemented. Then one transforms the slopes to loadings (λ) using equation (13), because the computational procedures for rotation have historically been worked out for factor loadings. Finally, EFA MIRT results are most often presented as a set of factor loadings, and the estimates of the correlations among the factors, because consumers of EFA results generally expect to see such results in terms of those parameters.

An Example: "The Affect Adjective Check List"

THE DATA

The Affect Adjective Check List (AACL) (Zuckerman, 1980) involves 21 adjectives; the first 11 ("Afraid" through "Upset" as listed in Table 18.19) are called the "anxiety-plus" adjectives, and the final 10 words ("Calm" through "Steady" as listed in Table 18.19) are "anxiety-minus" adjectives. To collect the data analyzed here (Odum Institute, 1988b), the adjectives were framed with the instructions "Please indicate whether or not the adjective listed describes how you feel today, today beginning with the time you woke up this morning." Anxiety-plus words are scored 1 if checked, and anxiety-minus words are scored 1 if not checked.

The item response data are from the same Computer Administered Panel Survey (CAPS) that collected the data used for the "impulsivity" example earlier. For this illustration, we use the data for the academic years 1986 through 1988, when the AACL was included in the CAPS study. The sample size is $N = 290$.

This dataset invites analysis with an obvious two-factor CFA model, with one factor for the anxiety-plus adjectives and a second (correlated) factor for the anxiety-minus adjectives. However, in this context we use these data as a textbook example of how item EFA can work.

EXPLORATORY ITEM FACTOR ANALYSIS

Table 18.19 shows the factor loadings and their standard errors, interfactor correlations, and goodness-of-fit statistics for two- and three-factor models for the AACL data. The lower-left portion of Table 18.19 shows the values of −2loglikelihood for one-, two-, and three-factor models that can be used with various criteria to aid in a decision about how many factors are needed to fit the data. The likelihood ratio test of the improvement of fit of the two-factor model over a unidimensional model is $G^2(20) = 441.13$, $p < .0001$, and for the three-factor model over the two-factor model the test is $G^2(19) = 74.57$, $p < .0001$; this suggests three factors are required. The AIC criterion (Akaike, 1974) also suggests three factors; however, the BIC criterion (Schwarz, 1978) suggests two factors. (Both AIC and BIC correct −2loglikelihood with [different] penalties for using additional parameters; both are used by selecting the model with the smallest value.)

Examining the loadings for the two-factor solution in Table 18.19, we see that the EFA procedure has almost perfectly divided the items into the "anxiety-plus" set, with large and significant loadings on the first factor, and the "anxiety-minus" set, with large and significant loadings on the second factor. The single exception is that the anxiety-minus adjective "Calm" has its largest loading on the first (anxiety) factor, but it is also the only adjective with a substantial split loading. In the two-factor solution, the factors for the anxiety-plus and anxiety-minus adjectives are correlated only 0.38, suggesting that the 21-item set is far from unidimensional.

The three-factor solution produces a doublet factor for the final two adjectives, "Secure" and "Steady"; otherwise, it is effectively the same as the two-factor solution. This is relatively unsurprising; it suggests that the anxiety-plus adjectives measure a reasonably unidimensional aspect of individual differences (anxiety), while the anxiety-minus adjectives indicate positive states, which are more multidimensional.

Conclusion

After a long period of development largely in the context of educational measurement, in the past decade IRT has become standard methodology for the construction of psychological scales and questionnaires. In this chapter we have described, and illustrated with examples, IRT models and methods that we have found most useful; these include the 2PL and graded models, with occasional support from analyses with the nominal model, DIF analysis, and MIRT. We have eschewed encyclopedic breadth in favor of illustrative examples based on real psychological data. Others' book-length treatments of IRT will provide the interested reader with more varied entry points to the field; recent examples include the texts by Embretson and Reise (2000), de Ayala (2009), and DeMars (2010).

Space limitations combined with our focus on topics within IRT that are useful in the context of virtually any scale-development project have led to the omission of some topics that may nonetheless

Table 18.19 Factor Loadings and Their Standard Errors, Interfactor Correlations, and Goodness-of-Fit Statistics for Two- and Three-Factor Models for the AACL Data

	2 factors				3 factors					
Adjective	λ_1	**s.e.**	λ_2	**s.e.**	λ_1	**s.e.**	λ_2	**s.e.**	λ_3	**s.e.**
Afraid	**0.97**	**0.08**	−0.18	0.07	**0.89**	**0.06**	*−0.17*	*0.05*	*0.13*	*0.20*
Frightened	**0.97**	**0.07**	*0.04*	0.16	**0.85**	**0.12**	*−0.04*	*0.15*	*0.29*	*0.18*
Fearful	**0.95**	**0.11**	*0.01*	*0.18*	**0.84**	**0.11**	*−0.07*	*0.16*	*0.28*	*0.17*
Worrying	**0.94**	**0.07**	*−0.08*	*0.13*	**0.96**	**0.11**	*−0.03*	*0.20*	*−0.08*	*0.35*
Terrified	**0.93**	**0.09**	*0.04*	*0.21*	**0.93**	**0.14**	*0.05*	*0.28*	*0.00*	*0.31*
Nervous	**0.91**	**0.11**	*−0.10*	*0.15*	**0.95**	**0.11**	*−0.06*	*0.17*	*−0.10*	*0.29*
Tense	**0.88**	**0.09**	*−0.04*	*0.14*	**0.93**	**0.11**	*0.04*	*0.20*	*−0.16*	*0.36*
Desperate	**0.81**	**0.11**	*0.19*	*0.17*	**0.84**	**0.06**	*0.19*	*0.19*	*0.00*	*0.27*
Panicky	**0.80**	**0.11**	*0.12*	*0.15*	**0.83**	**0.15**	*0.15*	*0.19*	*−0.07*	*0.35*
Shaky	**0.75**	**0.16**	*0.13*	*0.18*	**0.71**	**0.18**	*0.05*	*0.19*	*0.17*	*0.26*
Upset	**0.73**	**0.14**	*0.08*	*0.16*	**0.74**	**0.16**	*0.08*	*0.19*	*−0.01*	*0.28*
Calm	**0.53**	**0.13**	0.35	0.14	**0.52**	**0.17**	*0.31*	*0.16*	*0.08*	*0.22*
Joyful	*−0.09*	*0.18*	**0.96**	**0.09**	*−0.06*	*0.20*	**0.91**	**0.11**	*0.09*	*0.22*
Cheerful	*−0.07*	*0.22*	**0.89**	**0.17**	*0.05*	*0.17*	**0.98**	**0.13**	*−0.20*	*0.17*
Loving	*−0.23*	*0.17*	**0.86**	**0.12**	*−0.26*	*0.17*	**0.76**	**0.12**	*0.20*	*0.21*
Happy	*0.15*	*0.13*	**0.86**	**0.08**	*0.17*	*0.18*	**0.83**	**0.12**	*0.07*	*0.17*
Pleasant	*0.20*	*0.16*	**0.76**	**0.12**	*0.18*	*0.19*	**0.67**	**0.19**	*0.18*	*0.27*
Contented	0.32	0.12	**0.67**	**0.10**	*0.26*	*0.15*	**0.55**	**0.16**	*0.27*	*0.18*
Thoughtful	*0.04*	*0.17*	**0.65**	**0.13**	*−0.08*	*0.20*	**0.45**	**0.18**	*0.44*	*0.24*
Secure	0.42	0.13	**0.52**	**0.12**	*0.20*	*0.16*	*0.22*	*0.20*	**0.71**	**0.22**
Steady	0.48	0.13	**0.47**	**0.12**	*0.30*	*0.19*	*0.19*	*0.19*	**0.62**	**0.28**
Correlations:	0.38				0.34					
					0.44	0.41				
# of factors:		1	2	3						
−2loglikelihood:		5058.51	4617.38	4542.81						
difference:			441.13	74.57						
AIC:		5142.51	4741.38	4704.81						
BIC:		5296.64	4968.91	5002.07						

For each model, the largest loading in each row is bold, and loadings that are not more than twice their standard errors are italic.

be of special interest for measurement applications in clinical psychology. One such omission has been any discussion of the three- and four-parameter logistic (3PL and 4PL) logistic models, which add parameters to the 2PL model to represent non-zero item endorsement for low-θ respondents and probabilities less than one of item endorsement by high-θ respondents. Reise and Waller (2003), Waller and Reise (2010), and Loken and Rulison (2010) have shown that the phenomena represented by additional parameters in these models may arise in the measurement of personality and psychopathology (see also Thissen & Steinberg, 2009).

The use of IRT forms the basis for computerized adaptive testing. Because IRT scale scores on the same scale can be computed for any set of items, the items administered to any particular respondent can be customized to provide more precise measurement using fewer questions. While computerized adaptive testing has been widely used in educational measurement and personnel selection, it has only recently begun to be used in the measurement of psychopathology. Recently, the PROMIS initiative of the National Institutes of Health has developed a number of IRT-calibrated item banks for use in health outcomes research (Cella, Riley, Stone, Rothrock, Reeve, Yount, et al., 2010), including scales for the measurement of depression, anxiety, and anger in adults (Pilkonis, Choi, Reise, Stover, Riley, & Cella, in press) and children (Irwin et al., 2010). Several comparisons involving measures of psychopathology have shown that computerized adaptive testing provides precise measurement with fewer items than fixed scales (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Gibbons, Feldman, Crane, Mugavero, Willig, Patrick, et al., 2011; Gibbons, Grochocinski, Weiss, Bhaumik, Kupfer, Stover, et al., 2008). A thorough description of computerized adaptive testing would require a chapter or a book to itself.

Future Directions

While unidimensional IRT has become standard methodology for the construction of psychological scales and questionnaires, and recent developments have made MIRT computationally feasible as a part of item analysis, the computation and interpretation of multidimensional scale scores remain challenges. This challenge creates a limitation in the use of IRT for scales measuring various aspects of psychopathology, which frequently involve several interrelated constructs. MIRT scoring is the subject of active psychometric research at this time, and we anticipate that fully functional MIRT item calibration (and even computerized adaptive testing) systems will be available in the near future.

Acknowledgments

This work was funded in part by the National Institutes of Health through the NIH Roadmap for Medical Research PROMIS initiative, Grants 1U01AR052181–01 and 2U01AR052181–06 from the National Institutes of Health.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Bartholomew, D. J., & Knott, M. (1999). Latent variable models and factor analysis (*2nd ed.) Kendall's Library of Statistics*, vol. 7. London: Arnold.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model (ETS RR- 81–20)*. Princeton, NJ: Educational Testing Service.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37, 29–51.

Bock, R. D. (1997a). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16, 21–33.

Bock, R. D. (1997b). The nominal categories model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–50). New York: Springer.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika*, 46, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.

Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics, Vol. 26: Psychometrics* (pp. 469–513). Amsterdam: North-Holland.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75, 581–612.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology*, 59, 173–194.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Cai, L., Yang, J. S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.

Camilli, G. (1994). Origin of the scaling constant d=1.7, in item response theory. *Journal of Educational and Behavioral Statistics*, 19, 293–295.

Carragher, N., Mewton, L., Slade, T., & Teesson, M. (2011). An item response analysis of the DSM-IV criteria for major depression: Findings from the Australian National Survey of Mental Health and Wellbeing. *Journal of Affective Disorders*, 130, 92–98.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D. J., Choi, S. W., Cook, K. F., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., & Hays, R. D. on behalf of the PROMIS Cooperative Group. (2010). Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*, 63, 1179–1194.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, 36, 462–494.

Chen, W.-H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. *Quality of Life Research*, 19, 125–136.

Cole, S. R., Kawachi, I., Maller, S. R., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *Journal of Clinical Epidemiology*, 53, 285–289.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential item functioning in a satisfaction scale. *Journal of Applied Psychology*, 85, 451–461.

Covic, T., Pallant, J. F., Conaghan, P. G., & Tennant, A. (2007). A longitudinal evaluation of the Center for Epidemiologic Studies scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. *Health and Quality of Life Outcomes*, 5, 41–48.

Crawford, C. B., & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35, 321–332.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental Status Examination. *Medical Care*, 44, S134–S142.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology*, 74, 912–921.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement*. San Diego, CA: Robert R. Knapp.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, 63, 65–74.

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D., Schumacher, J., Saag, M., Kitahata, M. M., & Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research*, *20*, 1349–1357.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4–19.

Gibbons, R. D., Grochocinski, V. J., Weiss, D. J., Bhaumik, D. K., Kupfer, D. J., Stover, A., et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59, 361–368.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423–436.

Graham, J. R. (2000). *MMPI–2: Assessing personality and psychopathology*. New York: Oxford University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84–96.

Hancock, T. D. (1999). *Differential trait and symptom functioning of MMPI–2 items in substance abusers and the restandardization sample: An item response theory approach*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Harmon, H. H. (1967). *Modern factor analysis*. Chicago: University of Chicago Press.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holzinger, K. J., & Swineford, F . (1937). The bi-factor method. *Psychometrika*, 2, 41–54.

Hu, L., & Bentler, P. M . (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.

Irwin, D., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., Yeatts, K., Varni, J., & DeWalt, D. A. (2012). PROMIS Pediatric Anger Scale: An item response theory analysis. *Quality of Life Research*, *21*, 697–706.

Irwin, D., Stucky, B. D., Thissen, D., DeWitt, E. M., Lai, J. S., Yeatts, K., Varni, J., & DeWalt, D. A. (2010). An item response analysis of the Pediatric PROMIS Anxiety and Depressive Symptoms Scales. *Quality of Life Research*, 19, 595–607.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31*, 165–178.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202.

Jöreskog, K. G., & Gruvaeus, G. (1967). *RMLFA, a computer program for restricted maximum likelihood factor analysis*. Research Bulletin RB-67–21, Educational Testing Service, Princeton, NJ.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, 55, 312–320.

Krueger, R. F., & Eaton, N. R. (2010). Personality traits and the classification of mental disorders: toward a more complete integration in DSM–5 and an empirical model of psychopathology. *Personality Disorders: Theory, Research, and Treatment*, 1, 97–118.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York: Wiley.

Li, Y., Bolt, D.M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30, 3–21.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63, 509–525.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs,* Whole No. 7.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517–548.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8, 453–461.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association*, 100, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713–732.

McLeod, L. D., Swygert, K., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum Associates.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105–118.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.

Odum Institute. (1988a). *Eysenck Personality Inventory Form A (CAPS-EYSENCK module)* [Data file and code book]. Retrieved from http://hdl.handle.net/1902.29/CAPS-EYSENCK

Odum Institute. (1988b). *Affect Adjective Check List (CAPS-ANXIETY module)* [Data file and code book]. Retrieved from http://hdl.handle.net/1902.29/CAPS-ANXIETY

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, 14, 50–59.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., on behalf of the PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): Depression, anxiety, and anger. *Assessment*, *18*, 263–283.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Reckase, M. D. (2009). *Multidimentional item response theory*. New York: Springer.

Reeve, B. B. (2000). *Item- and scale-level analysis of clinical and nonclinical sample responses to the MMPI–2 depression scales employing item response theory*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Reeve, B. B. (2003). Item response theory modeling in health outcomes measurement. *Expert Review of Pharmacoeconomics and Outcomes Research*, 3, 131–145.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544–559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19–31.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8, 164–184.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27–48.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). The interactive effect of personality, time of day, and caffeine: A test of the arousal model. *Journal of Experimental Psychology: General*, 109, 1–31.

Revicki, D. A., Chen, W-H., Harnam, N., Cook, K., Amtmann, D., Callahan, L. F., Jensen, M. P., & Keefe, F. J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain*, 146, 158–69.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47, 361–372.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17, 34, Part 2.

Samejima, F. (1997). Graded response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, 6, 255–270.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. In C. C. Clogg (Ed.),

Sociological methodology (Vol. 18, pp. 271–307). Washington, DC: American Sociological Association.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53–61.

Schwarz, G. E . (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, 25, E34–E46.

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Implusiveness Scale: An update and review. *Personality and Individual Differences*, 47, 385–395.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa, City, IA.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, 66, 341–349.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, 81, 332–342.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology, *Psychological Methods*, 1, 81–97.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A., & Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly, 51*, 148–180.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 201–214.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43–75). New York: Routledge.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds), *Test scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19, 39–49.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). London: Sage Publications.

Thissen, D., & Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 123–144). Washington, DC: American Psychological Association.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–449.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 147–173). Washington, DC: American Psychological Association.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *A user's guide for SCORIGHT Version 3.0* (ETS Technical Report RR-04–49). Princeton, NJ: Educational Testing Service.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Woods, C. M. (2006). Ramsay-curve item response theory to detect and correct for non-normal latent variables. *Psychological Methods*, 11, 253–270.

Woods, C. M. (2007). Ramsay-curve IRT for Likert-type data. *Applied Psychological Measurement*, 31, 195–212.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, 33, 102–117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71, 281–301.

Yang, F. M., & Jones, R. N. (2007). Center of Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology*, 60, 1195–1200.

Yung, Y. F., McLeod, L. D., & Thissen, D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 113–128.

Zuckerman, M. (1980). The development of an affect adjective check list for the measurement of anxiety. *Journal of Consulting Psychology*, 24, 457–462.

CHAPTER

19 Missing Data in Psychological Science

Patrick E. McKnight *and* Katherine M. McKnight

Abstract

The inevitability and importance of missing data ought to move researchers to prevent, treat, and report the condition. Unfortunately, despite great advances in the field, researchers tend to ignore missing data. We hypothesize that ignoring missing data stems from low interest, unavailable solutions, and higher priorities by most social scientists. Thus, we aimed to remedy those potential mechanisms by providing a clear demonstration of missing-data handling in three distinct data analysis scenarios (psychometric, longitudinal, and covariance models) using R. Each of these exemplar procedures comes with code and data allowing readers to replicate and extend our examples to their own data. By demonstrating the use of missing-data–handling techniques in a freely available statistical package (R), we hope to increase available options and reduce the researcher's burden for handling missing data in common social science data analytic scenarios.

Key Words: Missing data, demonstration, R psychometrics, longitudinal analyses, covariance models

Introduction

Missing data, like death and taxes, is an inevitable fact that all social scientists must eventually confront. Like death and taxes, it is difficult to look forward to that confrontation, yet failing to do so can lead to some unintended and not-so-desirable consequences. As social scientists, we tend to focus on failed randomization, limited generalizability, participant recruitment, and other obstacles that interfere with our research objectives, yet missing data stands as one of the most prevalent and pressing scientific problems. Articles and books focus on the missing-data problem in the social sciences, but our science remains somewhat plagued by it. Recent surveys throughout psychological science (e.g., Peugh & Enders, 2005) and the continued stream of didactic articles (e.g., Graham, 2009) suggest the problem may be even more endemic than we imagine. To be sure, psychological scientists are compelled to resolve missing-data problems—along with other methodological concerns—but when to act and what to do remain a mystery for most of us; even the most seasoned quantitative methodologists struggle with the problem of missing data. So we have asked ourselves, what can be done for those of us who are not seasoned data analysts to help address missing data in an optimal manner? That is, what can we do to reduce and/or eliminate the unwanted consequences of missing data on our science? In this chapter, we highlight what we believe to be a lack of sufficient focus on the problem of missing data, and we offer a solution to directly address data analytic methods for handling a variety of missing-data scenarios.

The Problem

We believe that social scientists are less than optimally involved with preventing, treating, and reporting missing-data problems for various reasons, which can be summarized under three categories: interest, availability, and priorities.

Interest

As we have argued elsewhere (see McKnight et al., 2007), it is in the best interest of scientists to attend to missing data and the associated problems. The interest we refer to here is the collective focus of those involved throughout the social science enterprise on the problem of missing data. At the educational level, few training programs devote sufficient time and energy to teaching future social scientists about the consequences of missing data and what to do about them. In our experience, research methods and data analysis courses rarely give more than a cursory nod to the problem of missing data, and that tends to focus on quick and often problematic statistical solutions (e.g., listwise deletion, etc.; see our discussion, McKnight et al., 2007). The lack of focus on missing data in training programs—in our view—reflects the lack of focus in the field, where reviewers, editors, and publishers may pay less-than-adequate attention to missing data and its consequences. Although some funding agencies now demand more attention, the vast majority of publication outlets remain uncommitted. A cursory look at the most prestigious social science journals will demonstrate the nature of the problem: in any given study, sample sizes shrink, often without explanation, with each data analysis; intent-to-treat analyses reign supreme; and unexplained changes in degrees of freedom frequently go unquestioned by authors, reviewers, and most readers for that matter. Why does this lack of attention to missing data exist? It cannot be due to a lack of information; there is a plethora of books, chapters, journal articles, and online resources detailing the nature of the problem of missing data and methodological approaches to handling a wide range of missing-data scenarios. Yet missing data continues to be regarded as an esoteric topic on the fringes of psychological science, yet to permeate the fabric of our methodology.

Availability

It may be that the apparent lack of interest in the content of missing data comes from the complexity of the problem and general lack of readily available software to help address missing-data problems. Although software to handle missing data exists, specialized software for the more complex statistical routines (e.g., multiple imputation) is too expensive in terms of upfront software costs, human capital (the cost of expertise), and/or time (e.g., bug-ridden, overly complicated, or multistaged software solutions that take more time than traditional analyses) to be regarded as practical and/or useful. Even general-use statistical packages can fall prey to these problems. Furthermore, methods for preventing missing data (which we describe in detail elsewhere; see McKnight et al., 2007) often require resources that can appear to be and often are prohibitive. Statistical software, experienced data analysts or programmers with missing-data expertise, and dedicated data-collection personnel add stress to waning research budgets and for some create an access problem (particularly to data analysis expertise). If treatment and prevention of missing-data problems remain an obstacle, there is little wonder why our collective focus is targeted elsewhere.

Priorities

Our priorities reflect our interests. The aforementioned low interest in missing data and presumably scarce resources to address the problem leads to lower prioritization for journal space, reviewer attention, educational initiatives, and software development, among other consequences.

The lack of interest, availability, and prioritization we've discussed with respect to missing data resembles the bear–fish habitat cycle highlighted in James Gleick's book on chaos theory (Gleick, 1987). When left alone, bears and fish live symbiotically; bears eat fish and fish need bears to restrict the population. If we disturb this dependent life cycle, life expectancies of both populations are disturbed, leading to an eventual extinction of both species. Left undisturbed, bears and fish live in harmony by balancing out the food chain. Without stretching the analogy too far, we suspect that a disturbance in any of the three foci we've discussed with respect to missing data—interest, availability, or prioritization—would lead to a substantial change in missing-data treatment, handling, and reporting options.

A Potential Solution

It will not be easy to rectify the issues regarding missing data in psychological science. However, we can perhaps disrupt the interest/availability/prioritization cycle by changing the availability of statistical software for handling missing data, which may in turn reduce the lack of interest in and prioritization of missing data in the social sciences. In this chapter, our aim is to demonstrate how to use R—a freely available, open-source, platform-independent statistical package—for treating missing data in several common research designs. In previous efforts, we have elaborated more on prevention and reporting of missing data (McKnight & McKnight, 2009;

McKnight, McKnight, Sidani, & Figueredo, 2007), with some emphasis on treatment (Figueredo, McKnight, McKnight, & Sidani, 2000), but here we focus on treatment alone. By demonstrating the use of R, we hope to disrupt the availability problem by demonstrating to a broad audience how to treat missing data statistically, rather than ignore it, using freely available software that is easy and flexible to use. Our intent is to show that these tools are easy enough to use and, in so doing, eliminate the need for specialized programming knowledge. We realize that not everyone is technically oriented, and that can be a barrier to the availability problem with respect to statistical software. In this chapter we strive to make our demonstration as accessible as possible; however, it is important to point out that R is not a "point and click" program; that is, it is not menu-driven or graphical user interface (GUI) enhanced. Those who are used to working in such software environments may find the need to write syntax offputting. However, we have found, in our experiences teaching statistics and an R summer school course, that people tend to learn how to use R for their analyses quite readily. Our hope is that this demonstration will not only help our readers to implement procedures for handling missing data effectively, but also will increase the number of researchers who attend to the problems of missing data, including in the communication and publication of results. Each effort to chip away at the obstacles of interest, availability, and prioritization will, we hope, result in a shift toward the elusive but worthwhile aspiration of eradicating these missing-data problems.

A Step-by-Step Procedural Guide to Handling Missing Data in Psychological Science Using R

Preliminary Steps

Step 1. Installing and setting up R: Go to the r-project website (http://www.r-project.org) and follow the easy installation instructions. Be sure to attend to the specifics for the platform you are using (Windows, Mac, Linux, etc.).

Step 2. Run the program: After installation, starting R depends on the platform you use. Windows and Mac users simply click on the R icon; others can start it at a system prompt by typing the letter "R," without the quotes and in uppercase just as shown.

Step 3. Install the necessary packages: To run the missing-data procedures demonstrated in this chapter, you will need to install a few "packages" to follow along with our step-by-step guide. There are four packages you will need to install, all of which are listed in the syntax below. To install these tools, use the following command at the R prompt, paying close attention to all punctuation (e.g., quotes, commas, etc.), or else you will receive error messages:

```
install.packages("mitools","psy","lme4","mice","sem")
```

Step 4. Install additional packages: The first author created a set of missing-data procedures for the purposes of this chapter that are required to follow the demonstration; they are available at http://mres.gmu.edu/MissingDataR. Go to this website and install these functions in R in one of two ways: (1) by copying the Mdchapter.R file directly from the website and running them with the following code:

```
source("MDchapter.R")
```

or (2) by installing them with this line:

```
source("http://mres.gmu.edu/MRES/R/MD
chapter.R")
```

Now you are ready to start using R for the missing-data demonstration to follow.

Reading Data into R

In our experience, many social scientists use menu-driven or GUI-enabled statistical software such as SPSS. Therefore, this demonstration assumes that either you have your data in SPSS format (i.e., a *.sav file) or you can easily convert it into that format. R comes with a built-in function to read SPSS files, but you need to tell R that you need to use those functions. Follow these steps to do so:

Step 5. Reading SPSS data into R: Use the following commands to tell R that your data are in SPSS format, again being careful to include all punctuation:

```
library(foreign)
mydata = read.spss(file.choose(),F,T)
```

The "file.choose()" function pulls up a window where you can search for your SPSS data file and the "F" and "T" that follow specify that the read.spss function not convert value labels into factors and to return a data frame, respectively.

After you successfully complete this step, you will see a prompt ">" like the one you saw to run R. The data used throughout this chapter were randomly generated by our code. When you run our

examples using our code, the examples should differ because the data differ. Thus, there is no need to use your own data initially, but we thought it might be useful for readers to know how to move forward with their own data once they become familiar with the steps detailed below.[1]

Step 6. Check your data structure: You are ready to start analyzing your data. However, first make sure your data were read into R correctly by checking the structure (str) with the following command:

```
str(mydata)
```

If the data structure meets your expectations (i.e., variables and values are appropriate), you are ready to move to the next section.

Step-by-Step Instructions for Three Different Scenarios of Missing Data

Despite the perception that every research project produces its own missing-data scenarios, the problems can be grouped into some common themes. Although there are many, we focus on three common missing-data scenarios for social scientists, representing different types of problems and approaches to handling the missing data.[2] The first theme we address is focused on the measurement model, characterized by missing data for individual items from our research measures (e.g., surveys or questionnaires). Missing data at the item level interferes with our ability to compute composite or sum scores for our instruments, for subscales as well as full-scale scores. Clinical research using a depression inventory in which participants skip some items, whether purposely or not, is a good example. The second and more perplexing common missing-data situation arises in longitudinal studies where participants are followed over time or measured repeatedly. These within-subjects or even mixed-effects (within and between or split plot) designs present missing-data problems for us because if an individual drops out (or is not permitted to continue) or misses one or more observation periods, the data are missing at the person (vs. item) level. The third and final scenario focuses on the growing problem of missing data in covariance models (structural equation models [SEM]) where data analysis involves covariance matrices rather than raw observed values.

These three scenarios represent three of the more prominent data analytic areas in psychological science—psychometrics, hierarchical linear models (HLM) (in our case, as used for repeated measures), and SEM. We have chosen these three focal areas because we believe that they are readily extended into others. If readers have the basic tools to handle missing data in these three scenarios, they have acquired skills that span a broad range of missing-data situations in the social sciences.

The three missing-data scenarios require slightly different data for the required data analyses. The methods for producing our illustrative data examples can be downloaded from http://mres.gmu.edu/MissingDataR; we provide a brief overview here. We generated three complete datasets—each with sufficient observations and variables to demonstrate the given data analysis (e.g., HLM). For each of these three complete datasets, we generated two additional versions—one version that deleted data cells according to a random procedure (i.e., missing completely at random [MCAR]) and another version that deleted cells based on the data value of an observed variable (i.e., missing at random [MAR]). Therefore, for each of the three missing-data scenarios/demonstrations, we have three datasets (complete, MCAR, and MAR). We generated the data using the following commands:

```
mydat = createItemData()
mydat2 = createLongData()
mydat3 = createSEMdata(Nf=3)
```

The three datasets were then altered to generate missing values as MCAR or MAR with the following code:

```
mydat.mcar = createMCAR(mydat)
mydat.mar = createMAR(mydat)
mydat2.mcar = createMCAR(mydat2)
mydat2.mar = createMAR(mydat2,1,2)
mydat3.mcar = createMCAR(mydat3)
mydat3.mar = createMAR(mydat3,1,8)
```

Three different scenarios (item-level, longitudinal, and latent variable models) with three different patterns of data (nonmissing, MCAR, and MAR) resulted in nine total datasets. Each dataset generated by the functions above is unique, so the results demonstrated below are not representative of all results generated by any subsequent analysis. We offered the data-generation procedures simply to facilitate the use of these functions and familiarize readers with the procedures for handling missing data.

ITEM-LEVEL MISSING DATA

Consider a situation where clinical data gathered from a large number of patients in a specialty clinic contain missing values at the item level on one of our scales measuring a key variable (e.g., motivation for treatment). Our primary research question pertains to the relationship of this key variable, as measured by the total score on our scale, with other relevant variables. Psychometric data for this scale indicate that it is internally consistent (as indicated by Cronbach's alpha).

In this example, patients may have omitted responses because they forgot to complete an item, found an item irrelevant to their situation, or failed to understand the item wording sufficiently to answer it. Regardless of the reason, we find ourselves with incomplete data, and our analytic efforts might be adversely affected if we do not address these missing items. Before we treat the problem, we find it best to determine whether the extent of the problem warrants treatment. Missing a few items for a few people in a large dataset generally does not warrant the need for missing-data treatment. Conversely, large swaths of missing data across items and individuals presents a problem that might not be treatable. Thus, a quick diagnosis serves us well. To evaluate the extent of missing data using R, we run a descriptive summary on the dataset (in this example, "mydat"):

```
summary(mydat)
```

The dataset in this demonstration contains the observations for 200 patients on 20 items that make up the treatment motivation scale. Remember that the values summarized in the tables and figures were generated at random; your results will differ. It is more important to see the process than compare the specific estimates if you run the examples. The full dataset includes observations for all the other variables to be correlated with this treatment motivation variable. Due to space constraints, we suppress the summary results for the entire dataset and present information for several items for illustrative purposes. The summary should look like this for a single variable (in this example, a single item from the scale):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1.00	2.00	2.00	2.48	3.00	4.00	23.00

A summary for multiple variables will look like this, where i2 through i5 are the names of the variables (in this case, items 2–5 of our scale):

	i2	i3	i4	i5
Min	1	0	1	1
1st Qu.	2	2	2	2
Median	2	2	2	2
Mean	2.48	2.43	2.46	2.45
3rd Qu.	3	3	3	3
Max	4	4	4	4
NA's	23	19	16	25

The summary provides basic descriptive statistics, including the number of "NA's" or missing values by variable. It is recommended that prior to running the focal data analyses, the analyst reviews descriptive statistics to ensure accurate data values and to check for missing data. In addition to reviewing the summary of descriptive statistics for the dataset, a graphical depiction of the missing data can be informative.

The following R function produces an image graph where missing values are shown as white spaces and observed values are black. The following example shows the code for plotting the missing data (MD) for the MCAR dataset, where item values were deleted at random to create the missing data:

```
plotMD(mydat.mcar)
```

With the resulting image, the high contrast allows you to see if the missing values cluster by individual or item or both (Fig. 19.1). The image graph proves useful in all three missing-data scenarios detailed in this chapter. We present the missing-data image graph for the MCAR dataset in Figure 19.1 and for both the MCAR and MAR datasets in Figure 19.2.

The following code allows us to plot the same missing-data graph for several datasets side by side, as depicted in Figure 19.2. The first line tells R to create one row with two columns so we may have two plots side by side. Then, we issue the plotMD() function twice followed by another line to reset R back to a single plot window. These four lines produce Figure 19.2.

```
par(mfrow=c(1,2))
plotMD(mydat.mcar,main="MCAR")
```

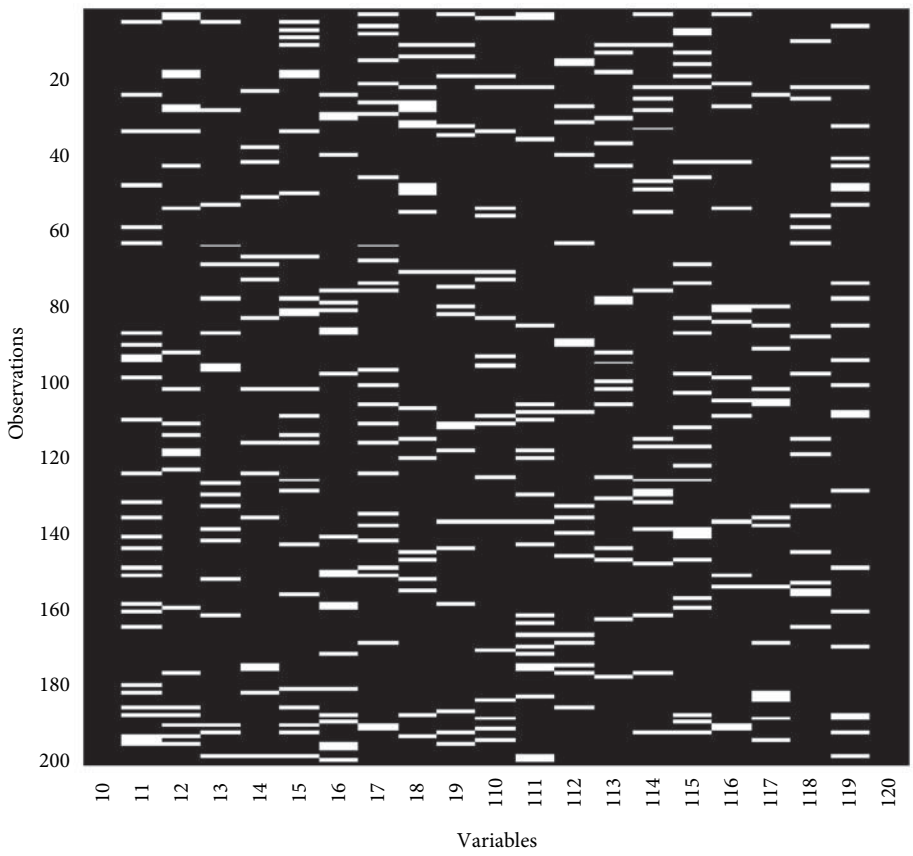


Figure 19.1 Missing-data graphical display using the plotMD() function in R, with the MCAR data.

```
plotMD(mydat.mar,main="MAR")
par(mfrow=c(1,1))
```

The plots show individuals by row on the *y*-axis (participants 1–200) and items 1 through 20 (I1–I20) on the *x*-axis. Reviewing the plots for both the MCAR and MAR datasets, it doesn't appear that any clear patterns emerge. For the MCAR dataset, there should be no discernable pattern because, again, the data were deleted randomly. For the MAR dataset, item values were deleted probabilistically, based on the observed values for item 1 (I1 on the *x*-axis of these plots). This suggests that a pattern does exist, but it is difficult to discern that pattern, which is often the case. When data are missing systematically, sometimes the pattern is easy to discern graphically, helping the analyst to better understand the nature of the missing data. For example, if the proportion of white spaces to black was high (i.e., a lot of missing data) and/or there was a large proportion of white spaces for a few items and they seemed to correspond with white spaces for other items, it may be that such a missing-data pattern is very difficult to handle statistically. Further analyses may be in order to ascertain the extent and pattern of missingness (see McKnight et al., 2007, for a detailed discussion on missing-data patterns and their meaning).

After analyzing the pattern of missingness, we conclude that moving forward with the focal analysis is warranted. One of the most efficient (statistically and time-wise) methods for treating item-level missing data when the scale is assumed to be unidimensional is by single imputation, whereby individual missing items get replaced by the mean item response. The missing-data literature refers to this procedure as "available item" analysis or "within-subjects" imputation. Many readers may be familiar with the problems inherent in mean imputation; however, the "mean imputation" method we discuss here is a within-subjects approach and not a between-subjects approach. That is, the mean is imputed for each person, versus groups of individuals. Because mean imputation is

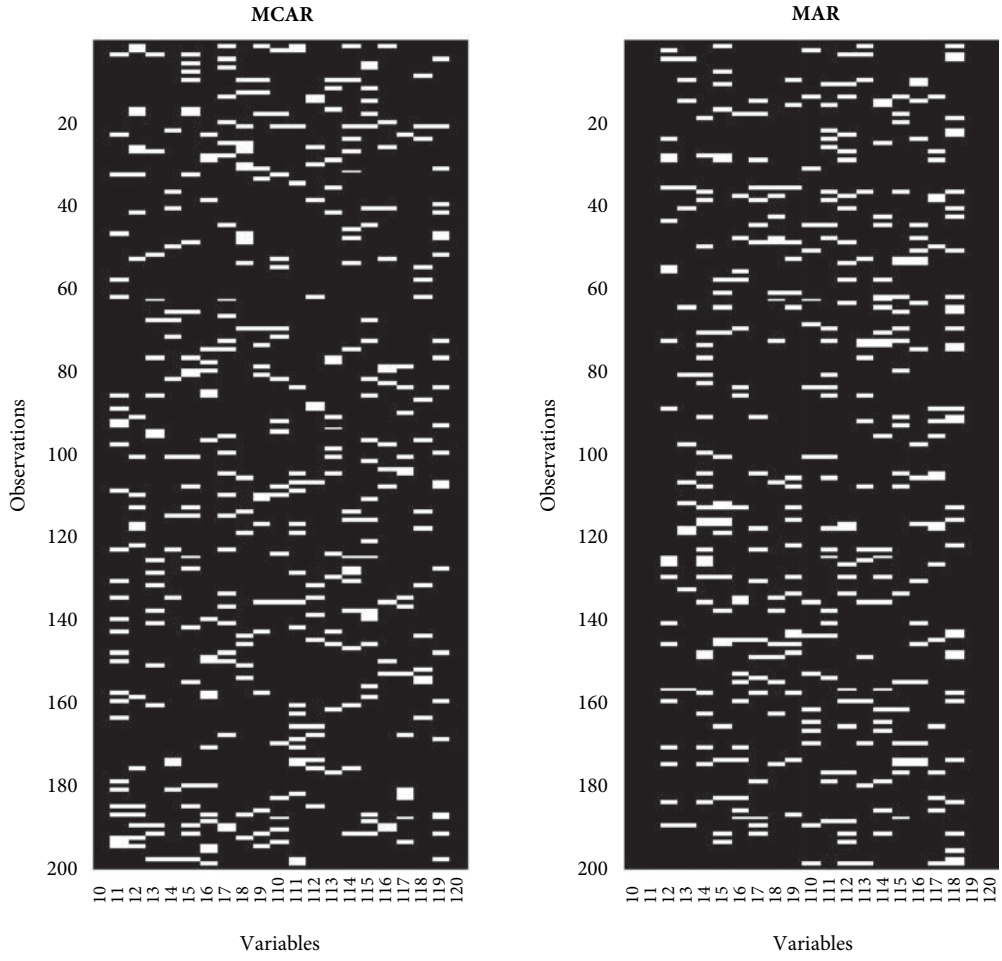


Figure 19.2 MCAR and MAR missing-data representation using the plotMD() function.

often a problematic method for handling missing data, it is critical to note the conditions under which it is useful and when it is not (see McKnight et al., 2007). In this example, where data are missing at the item level, the within-subjects imputation procedure is a strong option for handling missing data, provided that the data meet the following conditions: (1) the scale is known to be unidimensional—that is, items maintain a high level of internal consistency (alpha > .7); (2) the research question driving the data analysis focuses on the scale-level data (i.e., a scale score vs. item-level information); and (3) the scale-level scores are not inherently meaningful.[3] Most psychometric studies using classical test theory conform to these specifications. The first criterion rests on empirical verification, whereas the latter two rely on the selected or focal data analytic procedure. To ensure our data conform to the first criterion, we can perform some simple classical test theory analyses in R using Cronbach's alpha, the protocols for which are found in the library "psy" (again, replacing "mydat" with the name of your dataset):

```
library(psy)
cronbach(mydat)
```

The above code works only if the proportion of complete cases is large enough to compute alpha. In many cases, missing-item values across the sample reduce the number of available cases to the point where alpha can no longer be computed. Assuming you can compute alpha, you can then verify the internal consistency of your items from the data; otherwise, you must rely on the extant literature for estimates of alpha before moving forward. Table 19.1 shows somewhat comparable results for the different data, but the sample sizes differ greatly, reduced from 200 participants in the complete dataset to 32 in the MCAR dataset (where item values were randomly deleted to create the missing data), a reduction of 84 percent! However, if alpha is sufficiently

Table 19.1 Cronbach's Alpha Summary for the Three Datasets

	N	Items	Alpha
Complete	200.00	20.00	0.77
MCAR	32.00	20.00	0.84
MAR	75.00	20.00	0.77

high (at or above .70)—either based upon your data or the literature—then available item imputation or a within-subjects procedure remains useful.

Within-subjects imputation can be done by simply taking the mean of available items by person (as indicated by "row" in the following R code). The following code shows how to carry out within-subject imputation for replacing missing data by item with the individual's mean item score for a given measure. The function rowMeans() allows you to take the mean of all specified variables in your dataset and compute the mean across items. A total score (mydat$TotalScore) requires a single command in R:

```
mydat$TotalScore = rowMeans(mydat,na.rm=T)
```

Once computed, the new variable serves as either a predictor or outcome for subsequent analyses. All psychometric theories require items to be intersubstitutable (i.e., internally consistent) and related to a single underlying construct. Once the total scores are computed via this method, all observations contain a total score, and thus you have complete data for subsequent analyses.

To recap, missing data in scale scoring can be easily done by (1) assessing the extent of missing data either via numerical/tabular or graphical methods, (2) ensuring sufficiently high internal consistency by computing alpha with your data or collecting estimates from the literature, and (3) calculating a total score by calculating a mean for all nonmissing items per person.

MISSING DATA IN LONGITUDINAL OR REPEATED-MEASURES STUDIES

Most social science research requires a slightly more complicated procedure for handling missing data compared to the previous simple item-level problem. One situation that is certainly more complicated involves repeated-measures data where data may be missing for both predictors and outcomes. In those cases, we must use more sophisticated procedures for handling missing data. The following scenario provides a common situation where data are missing for both predictor and outcome variables, where the predictors represent between-subjects data and the outcomes represent within-subjects data. Our intention is to analyze the data in a linear mixed-effects model (often referred to as HLM in psychological science). Many readers may be familiar with the HLM approach and know that most statistical software packages estimate mixed-effects parameters using some form of maximum likelihood (ML)—a procedure known to produce relatively stable and efficient parameter estimates when data are missing. Despite ML's efficiency, we assume for the sake of this demonstration that the extent of missing data is such that our observed statistical power suffers greatly by using model-based procedures such as ML to treat missing data. Unfortunately for social scientists, such a missing-data scenario is not uncommon and requires a different approach to handling missing data than the default procedure of letting ML handle it without our intervention. This missing-data scenario requires direct intervention on the part of the data analyst and therefore provides an opportunity to demonstrate a simple application of multiple imputation.

Multiple imputation is the procedure of choice for handling missing data in situations where the number of missing values limits our sample size and, as a result, decreases our statistical power because it replaces missing values and therefore preserves the sample size. Moreover, ML is restricted to normal distributions; models that do not conform to normal distributions (e.g., Poisson models, logistic regression, etc.) may not be suitable for ML as a procedure for handling missing data. The following assumes that we want to analyze a two-level HLM model. Our study data consist of five repeated measures and two predictors (described below). We begin just as we did with the previous example by examining a simple summary of the data (the results of which are not shown here due to space constraints). Again, your summary results and figures produced by the following code will differ from ours below because each dataset is unique due to random number generation—both random values and random deletion methods.

```
summary(mydat2)
```

As before, we supplement the information provided by descriptive statistics with a graphical review of the missing-data situation by plotting the data using our plotMD function, as shown

in Figure 19.3. Again, for our demonstration, we plot the MCAR and MAR datasets side by side. The dataset contains five repeated measures (t1–t5) assessing level of substance use for 100 study participants, along with a baseline measure (b0), and two predictors—a continuous covariate (X1) assessing treatment motivation, and a two-level factor (F1) comparing a treatment to waitlist group.

Before we demonstrate the use of multiple imputation, we need to highlight some differences between cross-sectional and longitudinal data. Repeated observations may be organized in two different ways—a "wide" format, where the repeated measures are stored as additional variables, and a "long" format, where repeated measures are stored as additional rows for each subject. In this example, each of the five observations per person would be represented in a different row. Singer (1998) provides an excellent description of these different data structures. The data structure matters when imputation approaches are involved. Data in the wide format allow the algorithms to use all data by subject for imputation, whereas data in the long format restrict the imputation process to only data that fall within the specific time period. Consider the situation for our first observation in the MCAR dataset represented in Figure 19.3 where the participant failed to provide data for t1 but all other data were observed. If the data were organized in the wide format as depicted in the figure, then data from all time points (t2–t5) as well as the baseline variable (b0) and predictors (X1 and F1) would be used to compute the conditional probability distribution in a multiple imputation procedure using Multiple Imputation with Chained Equations (MICE). If the data were organized in the long format, then variables t2 through t5 are no longer available to conditionally impute values

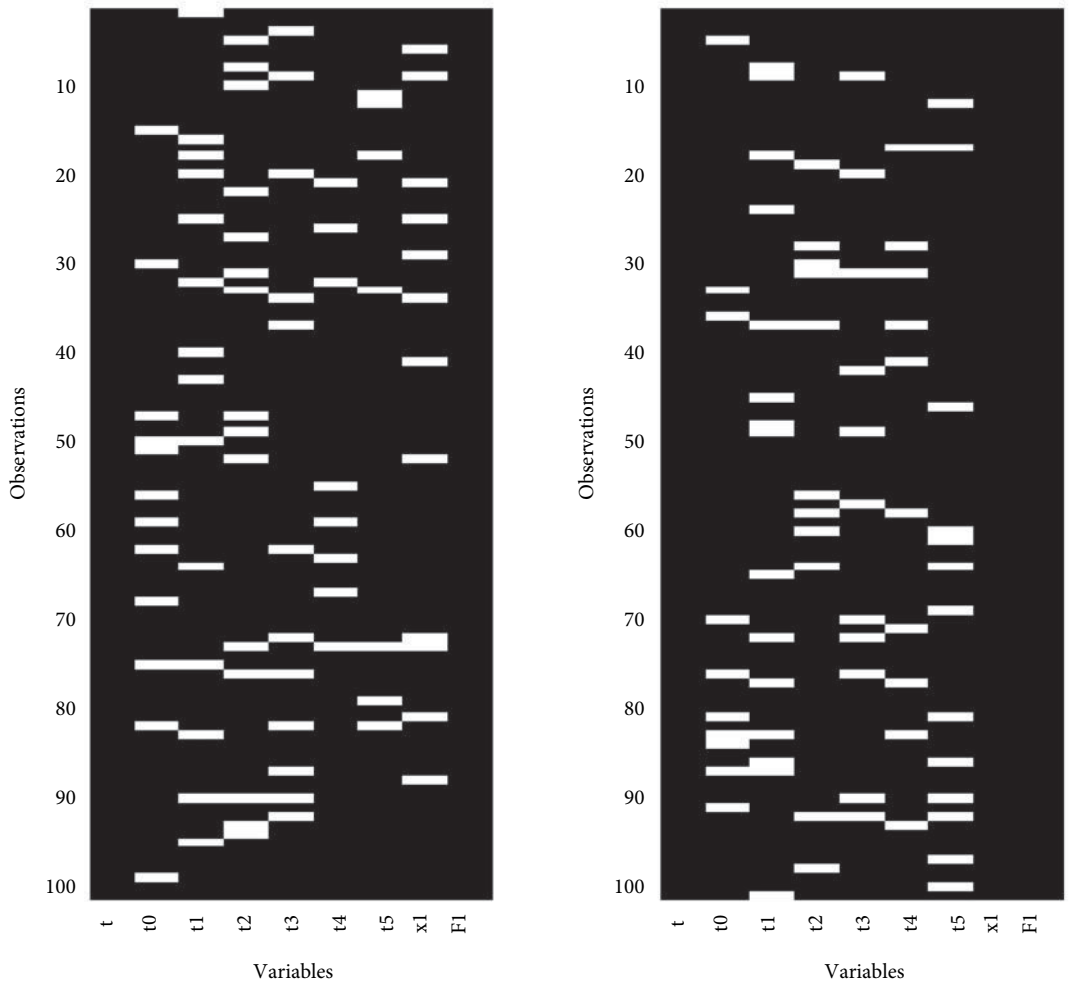


Figure 19.3 MCAR and MAR missing-data representation for the wide data.

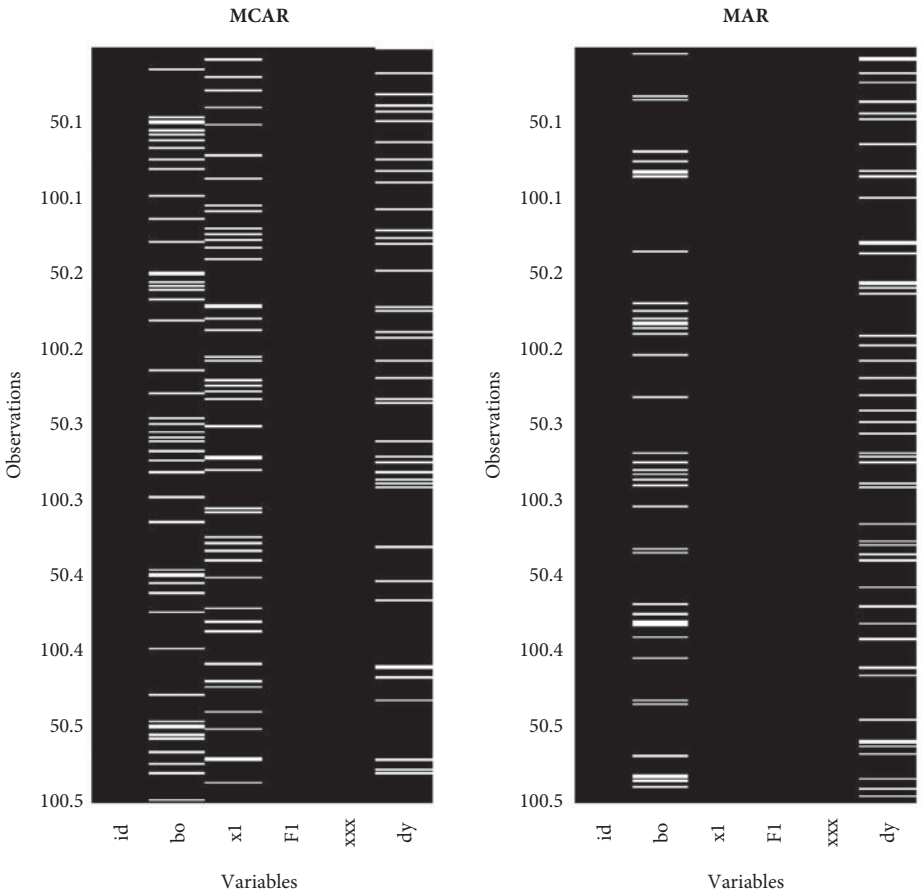


Figure 19.4 MCAR and MAR graphical missing-data representation for the long format.

for the missing t1 value because they appear as if they were different observations. Figure 19.4 shows why this is the case. On the *y*-axis, rows show that data for t1 for participants 90 through 100 occur (90.1–100.1), followed by rows for the same participants at t2 (90.2–100.2), and so on. Data for the two predictors X1 and F1 repeat row by row since they are time-invariant; that is, values for treatment motivation at baseline and the treatment group assignment remain the same throughout the five observation periods. The right-most column, "dv," is the individual's score on the dependent variable that is being measured at each of the five time periods and is therefore time-varying (in this example, dv = level of substance use). Only the data that appear across the row can be used for multiple imputation, and therefore t2 through t5 values cannot be used to impute missing values for t1, and so on. This is problematic in that we know that within-person repeated measures will be correlated and therefore useful for informing the imputation of missing values for those measures.

For these reasons and the fact that most social scientists are more familiar with and use the wide format for data management with repeated/longitudinal measures, our demonstration of multiple imputation is carried out on the wide format for this dataset.

The analytic models for the longitudinal analysis consist of a mixed-effects model using the following R code for a general linear mixed-effects model (lmer) model4:

```
lmer(dv~time+X1+F1+(1|id)+(time|id),
   data=mydat2.mcar.l)
```

where dv represents the value observed at each of the five time points (t1–t5); time represents the occasion the dv was measured (1–5); X1 is the continuous covariate (treatment motivation at baseline); F1 is the categorical binary predictor (treatment vs. waitlist group); (1|id) specifies a random intercept coefficient for each subject; and (time|id) specifies a random slope coefficient for each subject. Thus, the mixed model specifies fixed effects for X1 and

F1 but random growth parameters.[4] We do not contend that this is the proper statistical model for the given research question; our interest is simply to demonstrate the use of multiple imputation.

Step 1. Run MICE: The first step in the process is to run the multiple imputations. MICE is a multiple imputation procedure; the imputations are based upon conditional equations derived from the observed data. We first load the mice library using:

```
library(mice)
```

and then run the MICE function on our two incomplete datasets:

```
mydat2.mcar.mice = mice(mydat2.mcar)
mydat2.mar.mice = mice(mydat2.mar)
```

Both procedures above produce objects that contain five complete datasets. Multiple imputation replaces missing data with imputed values multiple times (thus the name) to increase the stability of the parameter estimates in the presence of missing data and to estimate the effect of the missing data on parameter estimation. The number of imputations is specified by the user, based on logic.[5] Users can change the number of multiply imputed datasets by adding to the R code above and specifying a new value for "m" (i.e., the number of multiple imputations) as modeled in the code below. Suppose, for example, you wanted to assess the stability of the parameters with 10 imputations. Y, you would change the syntax by adding an "m" value as we specify in the following code:

```
exampleWith10 = mice(mydat2.mcar,m=10)
```

We will continue our demonstration with the default value of m = 5; readers interested in more information about the MICE procedure are encouraged to read the mice() function help file in R or the published papers available online at http://www.multiple-imputation.com/.

Step 2. Reshape data: The second step in the process is to reshape the multiply imputed datasets that are in wide format to long format so they are suitable for the third analysis step. Reshaping data can be one of the most frustrating and complex procedures in R. Once you understand the process of reshaping, the function gets much easier to implement for a wider array of data frame structures. Our dataset was purposely structured to facilitate the demonstration; as a result, our reshape function was quite simple. Here is the code we ran to reshape each of the five datasets created by the mice() function:

```
for (i in 1:mydat2.mcar.mice$m){
assign(paste("mcar",i,sep="."),reshape(complete(m
ydat2.mcar.mice,i),varying=names(mydat2.mcar)
[3:7],idvar= "id",v.name= "dv", direction= "long"))
}
```

This simple "for" loop reshapes each of the wide complete datasets produced by the mice() procedure in Step 1 above. Five different long datasets are stored with the object names mcar.1 through mcar.5 via the assign() function above. The reshape function simply takes the wide format and reshapes the third through seventh variables (i.e., t1–t5, respectively) as a single variable "dv" and stores the suffix (1–5) as a new variable called "time." If you choose to run the mice() function on a long dataset, which can be problematic for the aforementioned reasons, then step 2 would be eliminated. Note that the reshape code listed above is for the MCAR data. If you would like to reshape the MAR data, the code would be as follows:

```
for (i in 1:mydat2.mar.mice$m){
assign(paste("mar",i,sep="."),reshape(complete(m
ydat2.mar.mice,i),varying=names(mydat2.mar)
[3:7],idvar= "id",v.name= "dv", direction= "long"))
}
```

Step 3. Set up imputation list: Before we can run the HLM or mixed-effects models, we must set up the data for easy secondary analysis. That setup requires us to specify the data to be analyzed. Remember that step 2 above provided us with five new completely observed, long-formatted datasets. We now must store those datasets into an imputation list using the following code, so that when we run the mixed-effects analysis, the same model will be run on these five separate datasets. The list enables us to work on a single R object with a function that expects that type of data structure.

```
library(mitools)
mydat2.mcar.list = imputationList(list(mcar.1,mc
  ar.2,mcar.3,mcar.4,mcar.5))
```

Similarly, the MAR data may be set up using this code:

```
mydat2.mar.list = imputationList(list(mar.1,mar.2,
  mar.3,mar.4,mar.5))
```

Step 4. Run LMER: After storing the multiply imputed datasets into an object that mitools can

Table 19.2 Multiple Imputation lmer() Results for the MCAR Dataset

	results	Se	(lower	upper)	missInfo
(Intercept)	0.25	0.08	0.10	0.41	4 %
time	0.23	0.03	0.18	0.28	7 %
X1	–0.02	0.03	–0.07	0.03	16 %
F12	–0.72	0.05	–0.82	–0.63	13 %

understand, we can now analyze the data using the lmer() function for mixed-effects models with the following code:

```
mcar.results = with(mydat2.mcar.list,lmer(dv~tim
  e+X1+F1+(1|id)+(time|id)))
```

or

```
mar.results = with(mydat2.mar.list,lmer(dv~time+X1
  +F1+(1|id)+(time|id)))
```

The code above produces lmer() results for each of the datasets listed in the imputation list specified in step 3 above.

Step 5. Summarize LMER results: The final and most important step in the multiple imputation process is to summarize the results for the statistical models. A multiple imputation summary is similar to any other model summary; however, instead of summarizing a single dataset result, R's mitools protocol helps us summarize all five completely imputed datasets by using the steps specified in the following code:[6]

```
mcar.betas = MIextract(mcar.results,fun=fixef)
mcar.vars = MIextract(mcar.results,fun=vcov)
mcar.vars2 = list()
for (i in 1: mydat2.mcar.mice$m){
mcar.vars2[[i]] = as.matrix(mcar.vars[[i]])
}
my.mcar.res = MIcombine(mcar.betas,mcar.vars2)
summary(my.mcar.res)
```

Summaries of the results are displayed in Tables 19.2 and 19.3 for the MCAR and MAR data, respectively. Note that only the fixed effects are summarized. The first column is the mean parameter estimate (i.e., mean betas pooled across the five imputed datasets) for the four fixed coefficients—the intercept, the slope (referred to as "time"), and the two predictors—specified in the lmer() function along with the standard error of those coefficients (column 2). Upper and lower 95 percent confidence interval bounds allow the data analyst to assess statistical significance; if "0" falls within those bounds, the fixed effect is not statistically significantly different than 0. Finally, the last column indicates the extent to which the missing data influence the parameter estimates. Higher values indicate that the imputations had a greater impact on the parameters and, as a result, indicate more missing information in the data as a result of missing data.[7] The MCAR and MAR results differ substantially given the fact that the amount and mechanism of missing data differed greatly between the two datasets. Missing iInformation values below 10 percent are probably not worth much attention, but values as high as 25 percent—as observed in the MAR results—ought to raise concerns us about the stability and replicability of these results, especially if the parameter estimate confidence intervals include 0. An example of that situation can be found in the MCAR results for X1, where the –0.02 parameter estimate has a 16 percent rate of missing information and the confidence intervals indicate a nonsignificant effect.

Table 19.3 Multiple Imputation lmer() Results for the MAR Dataset

	Results	Se	(lower	upper)	missInfo
(Intercept)	0.22	0.09	0.05	0.39	20%
time	0.23	0.03	0.18	0.29	15%
X1	0.02	0.02	–0.03	0.06	4%
F12	–0.74	0.05	–0.85	–0.63	25%

The five steps outlined above demonstrate the analysis of incomplete data using multiple imputation in linear mixed-effects models. As stated before, one advantage of multiple imputation is the retention of the study's sample size by replacing missing values based on observed values within the dataset. The result is multiple complete datasets with which to run the focal statistical models, improving the reliability of our parameter estimates and allowing us to estimate the impact of the missing data on those parameters. However, simply because we create complete datasets for analysis does not mean the generated results are valid. Through no fault of the statistics, complete case results might lead to erroneous conclusions since the extent of missing information produced by the missing data might generate nonreplicable effects.

MISSING DATA IN COVARIANCE MODELS

Our third and final demonstrated data analytic approach is the latent covariance model, for use when data analysis involves covariance matrices rather than raw observed values. Missing data in covariance models garners a fair bit of attention because many current SEM software packages such as Mplus (Muthén & Muthén, 2010) and AMOS (Arbuckle, 2006) offer procedures for handling missing data—including multiple imputation. Despite the availability of multiple imputation in these other software packages, we demonstrate how to run these models in R using code similar to what we used for the linear mixed-effects models in the previous section.

Our example model begins with as manifest variables (represented by boxes, as in Fig. 19.5) that are caused by three separate but related latent variables (F1, F2, and F3, represented as ovals). In other words, the manifest variables are indicators of the underlying, unobservable latent variables, and there are four indicators for each of those latent constructs. The nature of the relationship is a simple mediation model where F2 mediates the relationship between F1 and F3.

A plot of the missing data—for both the MCAR and MAR versions—is produced in Figure 19.6. Obvious from the two plots is the random nature of the MCAR missing values versus the clear pattern of missing observations located only in the upper half of the MAR dataset. This odd pattern is due to deleting values conditioned on a single variable (F3V4) in the dataset. We chose this data-deletion procedure to avoid computation problems that would make this demonstration too complex. Although the resulting missing-data scenario is generally unlikely, it enables us to provide a clean demonstration that is easily managed within the space constraints of this chapter.

A typical SEM in R provides all of the useful model fit indices and parameters that other packages provide. What separates R from other statistical software, however, is the common set of functions available for all analytic procedures.

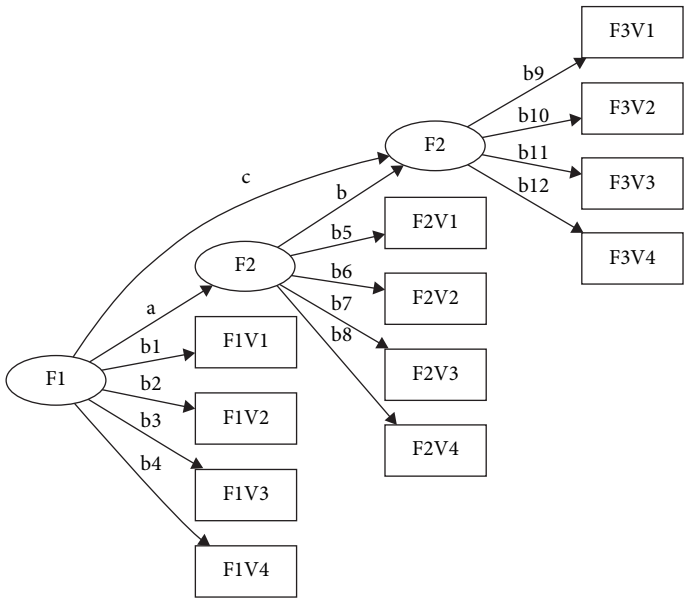


Figure 19.5 Demonstration model for SEM analysis using multiple imputation.

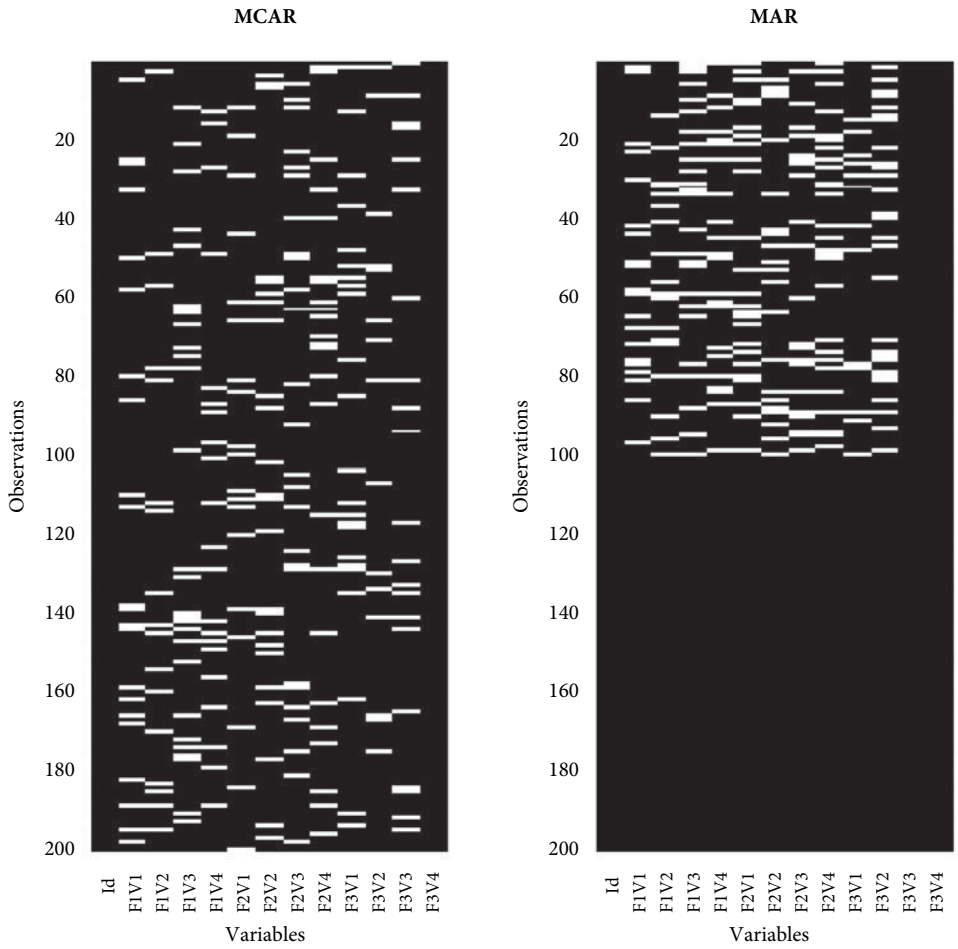


Figure 19.6 MCAR and MAR plotMD() for the SEM example data.

Analyzing an SEM model begins the same as the linear mixed-effects models using lmer(), except we need not worry about the data structure since the data are in the proper format necessary for our analysis. Thus, we provide the same step-by-step approach as before with a slight change. Some of the functionality for multiple imputation with SEM models requires additional coding. To eliminate that requirement for our readers, we provide the necessary functions just as we did for the patterns used to plot missing data using our plotMD() function. Below are the steps necessary to run the model depicted in Figure 19.5.

Step 1. Create SEM model: Fox and colleagues contributed a fully functional SEM package to the R community, and that package continues to be updated routinely by Dr. Fox and others (Fox, Kramer, & Friendly, 2010). The SEM package uses the reticular action model (RAM; McArdle & McDonald, 1984) format to specify the model. We encourage interested readers to consult the original RAM article along with Fox's detailed help file for the SEM package. The model depicted in Figure 19.5 is written in the following manner using the following RAM specifications:

```
SEMmod <- matrix(c('F1 -> F1V1','b1',NA,
'F1 -> F1V2','b2',NA,
'F1 -> F1V3','b3',NA,
'F1 -> F1V4','b4',NA,
'F2 -> F2V1','b5',NA,
'F2 -> F2V2','b6',NA,
'F2 -> F2V3','b7',NA,
'F2 -> F2V4','b8',NA,
'F3 -> F3V1','b9',NA,
'F3 -> F3V2','b10',NA,
'F3 -> F3V3','b11',NA,
'F3 -> F3V4','b12',NA,
```

```
'F1 -> F3','c',NA,
'F1 -> F2','a',NA,
'F2 -> F3','b',NA,
'F1 <-> F1',NA,1,
'F2 <-> F2',NA,1,
'F3 <-> F3',NA,1,
'F1V1 <-> F1V1','t1',NA,
'F1V2 <-> F1V2','t2',NA,
'F1V3 <-> F1V3','t3',NA,
'F1V4 <-> F1V4','t4',NA,
'F2V1 <-> F2V1','t5',NA,
'F2V2 <-> F2V2','t6',NA,
'F2V3 <-> F2V3','t7',NA,
'F2V4 <-> F2V4','t8',NA,
'F3V1 <-> F3V1','t9',NA,
'F3V2 <-> F3V2','t10',NA,
'F3V3 <-> F3V3','t11',NA,
'F3V4 <-> F3V4','t12',NA),ncol=3,byrow=T)
```

Each line represents the path between two variables. The paths labeled in our code as b1 through b12 are paths between a latent variable (e.g., F1) and a manifest variable (e.g., F1V1). Paths between latent variables (e.g., F1 to F2) are labeled as a, b, and c. Paths between any type of variable and itself (i.e., a variance parameter, such as F1V1<-> F1V1), are labeled t1 through t12. The long line of "NA" values at the end of each line specifies that the relationship has no starting value for the parameter and ought to be freely estimated. In three cases, in which the variance is specified for F1 to F3 (e.g., F1 <-> F1), the last value on the code line is 1. The value specified at the end of that line fixes the specified parameter to equal that value. In this example, we've fixed the variance of each latent variable, F1 to F3, to equal 1. Thus, we have a freely estimated model with only three fixed parameters—the variances for F1, F2, and F3. There are other, perhaps easier ways to specify the SEM model in R, but we chose to lay out the model in the most detailed way using the matrix syntax to avoid any confusion.

Step 2. Load Miruncombine.sem function: We created the Miruncombine.sem function to simplify the process of running multiple imputations and summarizing results within an SEM analysis in R, rather than having to write and run that code with each analysis. Miruncombine.sem uses the mice() function to produce multiply imputed datasets and the utilities provided in the mitools package to summarize the multiply imputed results. To load the function, simply install the MRES package or download the Miruncombine.sem function from http://mres.gmu.edu/MissingDataR and load the code with the following line, copying all syntax, including quotes and parentheses:

```
source("MDchapter.R")
```

Step 3. Run Miruncombine.sem function: Running the Miruncombine.sem function allows the analyst to run the multiple imputation process for treating the missing data and to analyze the SEM model created in step 1 simultaneously. To run the analyses, first assign an object name (e.g., sem.mcar is an arbitrary name given to our first SEM model using the MCAR data) to the Miruncombine.sem() function. Next specify the dataset you will be using (e.g., mydat3.mcar), then the dataset column that contains the variable representing the observation identifier (e.g., "id"). Then specify the structural equations model saved from step 1 (see above) and the number of multiple imputations (i.e., m) for the missing-data procedure. Your R code should look something like the following (here we've specified it for the MCAR [sem.mcar] and MAR [sem.mar] datasets for the SEM model depicted in Fig. 19.5):

```
sem.mcar=MIruncombine.sem(mydat3.
   mcar,idvar=1,SEMmod,m=5)
sem.mar = MIruncombine.sem(mydat3.
   mar,idvar=1,SEMmod,m=5)
```

The two objects, sem.mcar and sem.mar, store information that can be retrieved using the following code:

```
summary(sem.mcar)
or
summary(sem.mar)
```

We present the output/results in a cleaner manner than you will obtain from running the "summary" statement, for clarity. Tables 19.4 and 19.5 show the SEM results for both the MCAR and MAR data analyses.

Similar to the results from the lmer() models in the second example, the results from the multiple imputation procedure using the sem() function produce parameter estimates (i.e., mean unstandardized b's rather than betas for the SEM model), standard errors, confidence intervals, and missing-information estimates. The data in the left-most column refer to the tested paths from the SEM specified in step 1 and reflect the mean unstandardized

Table 19.4 Multiple Imputation Results Using the MIruncombine.sem Function on the MCAR Data

	Mean Parameters	se	(lower	upper)	missInfo
B1	2.29	0.13	2.04	2.55	0%
B2	2.17	0.14	1.91	2.44	10%
B3	2.11	0.14	1.84	2.38	3%
B4	2.29	0.14	2.02	2.55	2%
B5	1.99	0.12	1.74	2.23	7%
B6	1.85	0.12	1.61	2.08	5%
B7	2.00	0.14	1.73	2.27	17%
B8	1.91	0.13	1.66	2.15	3%
B9	1.59	0.12	1.37	1.82	9%
B10	1.85	0.13	1.59	2.11	23%
B11	1.70	0.11	1.48	1.91	4%
B12	1.85	0.11	1.63	2.08	0%
C	0.18	0.08	0.03	0.34	1%
A	0.28	0.08	0.13	0.43	1%
B	0.17	0.08	0.02	0.33	2%
T1	0.72	0.13	0.45	0.99	35%
T2	0.85	0.16	0.53	1.18	48%
T3	1.33	0.17	1.00	1.67	12%
T4	0.88	0.14	0.59	1.16	26%
T5	0.84	0.15	0.53	1.15	35%
T6	1.04	0.16	0.71	1.37	33%
T7	1.05	0.17	0.71	1.39	28%
T8	1.22	0.17	0.88	1.57	20%
T9	1.12	0.14	0.84	1.40	10%
T10	0.98	0.15	0.67	1.29	27%
T11	0.80	0.13	0.54	1.07	30%
T12	0.78	0.12	0.54	1.02	4%

parameter estimates across all five imputed datasets. Referring back to Figure 19.5, remember that b1 through b12 are the paths between the 12 manifest variables to one of the 3 latent variables (also known as the measurement model); a through care the paths between the latent variables (also known as the structural model); and t1 through t12 refer to the variances for each manifest variable. Path estimates are provided along with their standard errors, and the upper and lower limits based on 95 percent

Table 19.5 Multiple Imputation Results from the MIruncombine.sem Function on the MAR Data

	RMean Parameters	se	(lower	upper)	missInfo
B1	2.35	0.14	2.07	2.63	12%
B2	2.18	0.14	1.90	2.45	10%
B3	2.21	0.14	1.93	2.48	4%
B4	2.31	0.14	2.04	2.58	4%
B5	1.95	0.12	1.71	2.18	8%
B6	1.88	0.12	1.63	2.12	4%
B7	2.14	0.13	1.88	2.40	3%
B8	1.86	0.12	1.62	2.11	7%
B9	1.63	0.12	1.40	1.86	5%
B10	1.92	0.12	1.68	2.16	9%
B11	1.73	0.11	1.51	1.95	1%
B12	1.82	0.11	1.60	2.05	0%
C	0.19	0.08	0.03	0.35	0%
A	0.30	0.08	0.15	0.45	0%
B	0.18	0.08	0.03	0.33	2%
T1	0.79	0.16	0.46	1.11	46%
T2	0.97	0.14	0.69	1.25	20%
T3	1.18	0.18	0.83	1.53	30%
T4	0.91	0.14	0.63	1.19	20%
T5	0.71	0.12	0.47	0.96	24%
T6	1.15	0.16	0.84	1.46	18%
T7	0.94	0.15	0.63	1.25	22%
T8	1.08	0.15	0.78	1.38	22%
T9	1.18	0.17	0.84	1.52	30%
T10	0.83	0.15	0.53	1.13	30%
T11	0.94	0.13	0.68	1.19	6%
T12	0.89	0.13	0.64	1.14	1%

confidence intervals. Again, if those limits include 0, the path is statistically no different than 0, meaning the relationship between the two variables connected by the path is nonsignificant.

As noted in example 2 (our HLM model), one of the advantages of using multiple imputation for handling missing data is the ability to estimate the effect of the missing data on our statistical inferences, also called missing information or "missInfo" in the right column of these results. What is remarkable in this example is that despite the relatively large amounts of missing data, the structural parameters

estimated for the mediational model (i.e., parameters a, b, and c, the paths between the three latent variables) remain fairly unaffected by missing data. Both the MCAR and MAR data showed negligible variability in those parameters but huge variability in the parameters for the measurement models (paths b1–b12) and variance estimates (paths t1–12), giving us less confidence in those results. The instability in measurement model parameter estimates—across imputations—suggests weaker inferences in the measurement model but relatively strong inferences in the structural model.

Conclusion

Our aim has been to walk readers through three different data analytic scenarios to demonstrate the ease of handling missing data in R. We chose R because it provides a flexible and freely available platform for all data analysts. By sharing these demonstrations and missing-data packages with the end users, we hope to create a larger community of data analysts willing and able to treat missing-data problems. And with some additional optimism, we hope those users will produce, review, and perhaps oversee scientific contributions via editorships. Those involved in scientific inquiry need to be aware of the impact of missing data on study findings and be able to address missing-data problems via study design, statistical techniques, and reporting protocols, before our science becomes mired in unstable and nonreplicable findings. Our intention with this chapter is to add yet another plea to the literature imploring the field for more attention to data quality. By demonstrating the use of techniques for handling missing data in a freeware statistical package available to everyone, we hope to decrease the availability problem and reduce the burden on the end user for handling missing data in common social science data analytic scenarios.

Notes

1. The data used throughout this chapter come from artificially generated variables that can easily be generated by using the functions provided in the MRES package. Please consult the code referenced for this chapter for more details.

2. For a more in-depth discussion of missing-data scenarios, see McKnight et al., 2007, Chapter 3, in which we detail the more common scenarios based on a variety of facets aligned with Cattell's data cube.

3. In some research, the total scale score needs to be interpretable; for example, a measure of depression that provides critical cut-scores indicating either clinically and sub-clinically relevant depression then the scale score needs to be left in its original metric. We recognize that psychometric properties are not properties of the instrument per se but rather the properties of the instrument by respondent interaction. Nevertheless, we present the commonly held notion that these properties are attributable to the instrument to simplify the language and eliminate confusing elaborations where they serve little purpose for our discussion. We encourage readers unfamiliar with this distinction to consult Bruce Thompson's excellent article (Thompson & Vacha-Haase, 2000).

4. Here the reader may wish to refer to an introductory text on mixed effects models to understand the purpose of identifying which factors are considered fixed and which are considered random. Kreft and DeLeeuw's (1998) book provides a very readable introduction to multilevel models and these concepts. Another useful resource is Gelman and Hill's (2007) book on multilevel/hierarchical models

5. A detailed discussion at an introductory level about the multiple imputation process can be found in McKnight et al. (2007).

6. We omitted the necessary steps of testing multiple, nested models to ensure adequate error structures to simplify the demonstration. Those interested in a complete approach should consult Pinheiro and Bates (2000). The lmer() function comes in the lme4 package. You must run library(lme4) before running any commands using the lmer() function. Please consult the code we produced for this chapter at the following website: http://mres.gmu.edu/MissingDataR.

7. See McKnight et al. (2007) for an introductory-level discussion about the missing information parameter or gamma (γ).

References

Arbuckle, J. L. (2006). *Amos (Version 7.0) [computer program]*. Chicago: SPSS.

Figueredo, A. J., McKnight, P. E., McKnight, K. M., & Sidani, S. (2000). Multivariate modeling of missing data within and across assessment waves. *Addiction*, 95(suppl. 3), S361–S380. Fox, J., Kramer, A., & Friendly, M. (2010). *SEM: Structural equation models*. R package version 0.9–20. http://CRAN.R-project.org/package=sem0

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gleick, J. (1987). *Chaos: Making a new science*. New York: Viking.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.

Kreft, I., & DeLeeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

McArdle, J. J., & McDonald, R. P. (1984) Some algebraic properties of the reticular action model. *British Journal of Mathematical and Statistical Psychology*, 37, 234–251.

McKnight, K. M., & McKnight, P. E. (2009). Measures for improving measures. In D. L. Streiner & S. Sidani (Eds.), *When research goes off the rails* (pp. 268–279). New York: Guilford Press.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J . (2007). *Missing data: A gentle introduction.*New York: Guilford.

Muthén, L. K., & Muthén, B.O. (1998–2010). *Mplus user's guide (6th ed.).* Los Angeles, CA: Muthén & Muthén.

Peugh, J. L., & Enders, C. K. (2005). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer Verlag.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24(4), 323–355.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60, 174–195.

PART 4

Matters of Responsible Research Conduct in Clinical Psychology

This page intentionally left blank

CHAPTER

20 Ethical Considerations in Clinical Psychology Research

Gerald P. Koocher

Abstract

Clinical research requires careful attention to ethical principles from the design of the project, through data collection, and extending on to the data analysis and publication of results. Each of these stages requires attention to different issues in order to fully inform, respect, and protect research participants. In addition, special attention must focus on the integrity of the research enterprise and our relationships with co-investigators, assistants, and students. New technologies and new professional skills will offer increased opportunity, but will also require that we confront new challenges

Key Words: CABLES, clinical trials, research ethics, research regulation

Introduction

The simplest way to conceptualize the ethical problems associated with clinical research involves a three-step process following the sequence in which investigators typically undertake their work: (1) planning the study prior to engaging participants; (2) collecting the data, with engagement of participants; and (3) the analysis and dissemination of findings, which may or may not involve ethical obligations to the participants. This chapter follows that same sequence, raising ethical concerns during the stage of the research where they will most likely arise. Although the general principles apply to all research involving human participants, this chapter will focus chiefly on research in clinical psychology as contrasted with the broader spectrum of biomedical research.

In most cases researchers begin planning their studies long before any participants walk into the lab or interview room. The key exception occurs when the research project involves reanalyses of data already collected or the mining of data from archival sources. In all cases, however, this beginning stage involves establishing hypotheses and methods that one plans to deploy, securing appropriate authorizations, and identifying or recruiting potential participants. If done well, the planning work accomplished in this stage will protect participants going forward and ensure compliance with well-understood ethical and regulatory standards.

The next stage of the research venture involves data collection. At this point the investigator–participant interaction begins with recruitment, obtaining consent, and conducting the study. The research may take the form of a simple one-time questionnaire or a long-term project involving many sessions and different types of data collection or follow-ups. The project may involve mundane or highly sensitive issues. Participation may involve oral, written, or physiological measurements. Those asked to participate may come from any age range and can include people with significant psychological symptoms or those with no discernible psychopathology. We may offer some people the chance to enroll in a clinical trial of a new treatment that could benefit them. We may also seek participants

to learn about their lives or mental condition with no hope or promise that they will personally benefit from participating. In all of this work we must stand ready to provide all necessary ethical protections to those we study across this broad range.

The final stage of the research venture involves reaching valid conclusions related to our hypotheses from the data we have collected and disseminating the results. We must conduct our analyses and reporting with accuracy and integrity. We must also consider our obligations to colleagues as partners in the scientific venture when assigning authorship in publications. And we must consider how best to distribute our findings in ways that promote the science of psychology and human welfare.

In the Beginning

Planning the Project

The *CABLES* acronym and metaphor (Koocher, 2002) articulates a means of conceptualizing research participation risk by considering six distinct strands of risk that may pose harm to research participants may exist: cognitive, affective, biological, legal, economic, and social/cultural. A year later another publication (National Research Council, 2003) offered a remarkably similar set of descriptors, which it termed psychological, physical, legal, economic, social, and dignitary. By considering such dimensions when planning a research project, investigators can often anticipate and take steps to mitigate potential harms. For the sake of illustration, this chapter focuses on the original *CABLES* conceptualization. A summary of the CABLES categories and examples appears in Table 20.1.

Cognitive risks include threats to the participant's intellectual functioning, learning, academic achievement, and the thoughts underpinning self-esteem and emotions. This category covers all hazards to cognitive functioning that do not result from anatomical or biological changes. Examples of research in this category might include any type of clinical trial. Consider comparative analyses of psychotherapy treatments, teaching strategies, remediation of learning disabilities, or problem-solving tasks. Imagine a study of a new experimental high school mathematics curriculum that results in participants from the control group (i.e., enrolled in the standard math curriculum) earning higher or lower scores on the Scholastic Assessment Test. Were the enrolling students alerted to this important potential benefit or problem from the outset of their participation? What happens when a randomized clinical trial to test a new remediation strategy for a specific psychological problem proves very successful? Do the participants in the control group gain access to the more effective approach that they missed out on? What if the participants in the control group are offered the newly proven beneficial treatment subsequently, but have passed through a critical developmental period (e.g., completing the college application process based on the lower test score) that vitiates the benefits of the program (e.g., the control group members feel "dumber" than their peers, or their lower test scores are all that is on file as they apply for college admission)? Consider the participant challenged to solve difficult or insoluble problems in a school setting as part of a psychological experiment in frustration tolerance. What if the inability to succeed at the challenging task wounds the participant's self-esteem, leading to a loss of motivation that generalizes beyond the laboratory?

Affective risks are the hazards of emotional distress both during and following participation in the research. These might include risks of self-discovery when participation in research reveals aspects of oneself that the participant would rather not see, as in Milgram's (1963) classic study of obedience. In Milgram's deception paradigm many participants followed orders to administer what they thought were painful electric shocks to a confederate of the experimenter. Even after debriefing, some participants felt troubled by inflicted insights (i.e., "How could I have agreed to do such a thing? What kind of a person am I?").

Other emotional risks may occur in tests of exposure-based therapy techniques (Boudewyns & Shipley, 1983), where the intensity of the treatment may seem worse than the symptoms to some participants. Alternatively, participants in placebo control trials where the ineffectiveness of an intervention or placement on a waiting list offered to members of a symptomatic control group may lead them to experience continued distress or give up hope while enrolled in the study.

Emotional risks may also occur in studies that involve collection of sensitive data such as HIV infection or genetic risk status (Bayer, 1989; Burris & Gostin, 1997; Hudson, Rothenburg, Andrews, Kahn, & Collins, 1995; Lapham, Kozma, & Weiss, 1996; Moreno, Caplan, & Wolpe, 1998). Participants entering such studies may not be fully prepared to deal with the consequences of what they discover. Such information can also trigger legal (e.g., liability and discrimination claims) and economic hazards as well (e.g., loss of employment and inability to obtain insurance). In addition,

Table 20.1 CABLES Categories and Examples

CABLES Category	Description	Examples
Cognitive	Risks to intellectual functioning, learning, or other nonbiological hazards to cognitive abilities	• Assignment to a "no-treatment" or "waiting list" control group causes a learning-disabled child to miss out on a helpful tutoring intervention or a depressed adult to miss out on an innovative CBT treatment at a point in time when the intervention might have provided a critical benefit.
Affective	Risks of emotional distress, including postdebriefing distress in studies involving deception	• A group of college students undergo prescreening to assess eligibility for an intervention aimed at a socially undesirable trait (e.g., racism, social awkwardness, etc.). As a result of the screening they learn that they scored high on the trait and suffer damage to their self-esteem. • People who participate in a genetic screening program with alternative counseling options become severely depressed after learning of their carrier status, despite the intervention provided.
Biological	Risks of physical injury or illness during or as a result of participation	• Some participants in a clinical trial involving medication suffer unanticipated noxious side effects. • Biofeedback equipment in poor repair causes electrical burns to participants in a study. • Participants contract hepatitis following exposure to contaminated equipment while providing a blood specimen.
Legal	Exposure to legal hazards as a result of participation	• During a research interview the participant provides information that triggers a mandatory report to authorities. • Qualitative interview data collected for behavioral science research on offenders are subpoenaed by law enforcement authorities who learn of the study.
Economic	Costs associated with actual financial loss, lost opportunity costs, or costs of remediation of damages resulting from participation	• Research procedure requires a participant to take unanticipated time off work, causing a financial loss. • Due to a "misunderstanding," the participant incurs unreimbursed travel costs to the experimenter's lab.
Social/Cultural	Hazards related to social rejection or stigmatization as a result of participation	• Individuals or groups of participants acquire a stigmatizing label (e.g., "at risk," "alcoholic," or "underachiever") as a result of their participation. • Social or cultural groups stigmatize or reject individuals because of their participation in a particular project (e.g., presumed HIV seropositivity or genetic carrier status).

emotional distress may follow an effort to enroll in a potentially lifesaving clinical trial only to be screened out as ineligible, or considerable guilt-driven distress may follow if genetic testing reveals that one individual is unaffected by an ominous gene present in other family members.

Biological risks refer to the hazards of physical injury or illness as a result of delayed, ineffective, or absent treatment; as a direct or side effect of the intervention; or as a result of investigator negligence. The biomedical ethics literature is replete with stories of such hazards and late effects, ranging from the "medical experiments" conducted in Nazi concentration camps (Annas & Grodin, 1995; Spitz, 2005) to more recent disclosures of radiation experiments using terminally ill medical patients

and armed forces personnel conducted by or for the U.S. military and Department of Energy during the Cold War (Koocher & Keith-Spiegel, 2008). From a mental health perspective one could cite many psychopharmacology studies, such as the early investigative work with clozapine as a treatment for schizophrenia, in which several deaths occurred from drug-related agranulocytosis, a severe drop in white blood cell counts.

At a more simplistic level, researchers must take care to ensure the safety and maintenance of any equipment they use in their procedures (e.g., biofeedback equipment, electrodes, or apparatus of any sort). All lab assistants, phlebotomists, or others involved in data collection should have appropriate safety training. Protocols should specify follow-up procedures for any participant injuries—even slipping on the lab floor.

Legal risks might include adverse consequences, such as disclosure of sensitive identifiable confidential information, statutorily mandated reporting of abuse or neglect, or even insight-generated legal actions (i.e., litigation begun as the result of self-discoveries made in the course of one's participation in research). As noted earlier, a number of sensitive areas of medical research, including diagnosis and treatment of people infected with HIV/AIDS and predictive genetic testing, may lead to significant legal risks (Bayer, 1989). Psychosocial research in areas such as child neglect or abuse, child or caregiver substance abuse, and domestic violence can pose similar risk, especially if confidentiality and data access are inadequately protected (Wolf & Lo, 1999).

In one study, for example, an investigator sought to determine how children who had never been sexually abused played with anatomically detailed dolls and contrasted their responses with children who were known sexual abuse victims. In seeking participants for the control (i.e., nonabused) group from among large numbers of nursery school children, it was necessary to screen out children who may have been abused previously. Asking that exclusionary question directly of the parents of prospective participants might conceivably have yielded answers that would require the investigators to report to authorities previously undiscovered cases of suspected child sexual abuse. To avoid this risk, the researchers presented a list of multiple exclusion criteria (e.g., Do not agree to allow your child to participate if he or she has recently been exposed to an infectious disease, does not tolerate interacting with strangers well, is currently enrolled in psychotherapy, or may have been sexually abused in the past). When parents subsequently declined to enroll their children as participants, the investigators did not have any data that might trigger a required report under statutory child-abuse reporting mandates.

Economic risks are actual financial hazards associated with incurred costs (e.g., transportation to experimenter's laboratory) or lost opportunity costs (e.g., time and revenue lost from paid employment to enable participation), and remediation of damages associated with participation in the research. In one actual study of job interview skills using a deception paradigm for the sake of realism, the investigator advertised a highly desirable employment opportunity. Applicants came for interviews only to find out after the fact that there were no actual job openings; rather, they had been unwitting conscripts in a "naturalistic experiment." At least one such participant had purchased new clothes and taken unpaid time off from another job to attend the "job interview."

Social and cultural risks leading to social rejection or stigmatization are again well documented in studies of HIV/AIDS, substance abuse, or genetic risk factors (Bayer, 1989; Hudson et al., 1995; Lapham et al., 1996). At the same time, however, even research on relatively benign topics can lead to a degree of stigmatization, with some members of a group feeling excluded based on study criteria while friends or classmates are included. In some studies of "gifted" or "socially isolated" children, for example, it can be all too obvious to peers that it is "good" or "bad" to be chosen for a particular research group. The famous study of *Pygmalion in the Classroom* (Rosenthal & Jacobson, 1992) also illustrated this point well. In that study, teachers, who were led to expect positive changes in their students' academic performance based on bogus psychological test data, subsequently rated the students as having fulfilled the prophecy of the phony data. It appeared that the teachers altered their expectations and ultimately rated their students more positively, with beneficial results for some students. Suppose the teachers had been told the students were "predelinquent" or "potential school dropouts." Imagine the impact of such a stigma.

In other studies, certain participants deemed to be at risk for some type of problem or bad outcome are selected for study. Examples of at-risk populations include the children of schizophrenic parents, recently divorced people, preschoolers from disadvantaged homes, parents who fit patterns indicating a susceptibility to abusing their children, and people functioning under high stress. Educational,

psychotherapeutic, and coping and skill-building training are among the interventions frequently employed. However, such research often fails to consider the potential consequences of acquiring the "at-risk" label, especially when the participant deemed at risk becomes known as such to others (e.g., parents, teachers, or community members) and assigned to a control group or an arm of the study that offers no direct benefit or subsequent remediation.

All too often, investigators fail to consider special factors related to race, culture, or ethnicity that may complicate the validity of their research or adequately respect the participants. For example, historical mistrust of "outsider" investigators and ethically questionable research methods has led some groups and communities to become averse to participation in research (Darou, Hum, & Kurtness, 1993; Gamble, 1993; Harris, Gorelick, Samuels, & Bempong, 1996). In such situations the design of studies and the construction of control groups require particularly thoughtful consideration of cultural differences and historical abuses as well as careful community consultation (Fisher & Wallace, 2000; Norton & Manson, 1996).

By considering all strands of the CABLES model in the planning and design of research, investigators can anticipate and prevent many ethical challenges. The remainder of this chapter will focus in greater detail on the scope of topics to consider as an investigator moves more actively into the conduct of the research.

Competence to Conduct Research

The scientific merit of a research design has been widely acknowledged as a competence issue, and many mental health providers have not had extensive training in research design and data analysis. These techniques have become extremely sophisticated in the past few decades, largely made possible by the advent of the high-speed computer. Researchers who lack such skills should include someone with expertise in design and statistics on their research team, at least as a consultant. No meaningful information can possibly result from poorly formulated studies or improperly collected or analyzed data. The use of human beings or animals in research cannot be justified on any grounds if the study is flawed. At best, the participants' efforts are wasted, and, at worst, they could suffer harm.

The scientific record also becomes tarnished when poor-quality work is dumped into the scientific literature. Ideally, the editors of scholarly journals weed out most incompetent submissions, but shoddy work slips by for a variety of reasons, such as inadequate or biased reviews. The more critical the topic (e.g., heart disease as opposed to heartburn), the more willing manuscript reviewers may be to overlook or underrate methodological flaws (Wilson, DePaulo, Mook, & Klaaren, 1993).

Even the most proficient researchers face many serious dilemmas when designing their projects. Quality science and ethical research practices typically provide the best results, but scientific merit and ethical considerations are sometimes at odds, requiring the sacrifice of some measure of one to comply with the other. For example, fully informing the participants of the purpose of the study may weaken or distort scientific validity. The privacy of vulnerable people may be invaded in long-term follow-up studies designed to evaluate and improve treatment techniques from which they and others may eventually benefit. Balanced placebo designs may require misinforming participants to reduce the effects of expectancies. Participants in a control group could be denied a valuable experimental treatment, but without a control group that treatment could not conclusively be proven superior. Fortunately, many seeming conflicts between science and ethics can often be resolved or minimized by competent researchers after careful reflection, consultation, and reworking of the original plan (Sieber, 1993).

Regulatory Compliance and Institutional Approval

A large literature on the responsible conduct of social and behavioral research has developed over the past several decades. Interest surged shortly after World War II when the Nazis' obscene interpretation of what constituted legitimate science and the criminal acts they committed in the name of science became known (Annas & Grodin, 1995; Spitz, 2005). Concerns accelerated with revelations of questionable and risky procedures used on human beings without their voluntary and informed consent in other countries, including the United States. Consider, for example, the Tuskegee study, wherein poor, black, syphilitic men in Alabama were left largely untreated for the purpose of understanding how that ravaging disease progressed. As many as 400 men may have lingered and died from a curable disease (Jones, 1981). How such a study could have been publicly tolerated from 1932 until the early 1970s remains a matter of debate and consternation (Koocher & Keith-Spiegel, 2008).

Researchers must now adhere to high standards of care, many of which align with the duties owed patients by mental health service providers. However, clients in need of mental health care usually present themselves for services with the understanding that they will receive assessment or psychotherapeutic services. Typically, research participants must be actively sought out and do not always know or fully understand the nature of the research; sometimes they do not even recognize that they are under study. When offering therapeutic and assessment services, the practitioner focuses on meeting the client's needs. On the other hand, data collection constitutes the means by which researchers achieve their goals. In general, practitioners hold the interests and welfare of each individual as primary, whereas researchers contend with the motivation to also fulfill personal agendas that can, without constant self-monitoring, overshadow the rights and welfare of those they study.

Because of an evolving understanding of past abuses and the adverse consequences to some participants, the federal government began creating research policies in the 1950s. A complex web of regulatory bodies and policies now require researchers to become familiar with a number of acronyms and concepts that demand careful consideration, including IRBs, HIPAA, FERPA, DSMBs, informed consent, minimal risk, clinical equipoise, and therapeutic misconception. Thumbnail introductions to each of these key policies or concepts follow; however, the Center for Ethics Education at Fordham University provides a substantial compendium of resources and guides to both national and international standards and regulations, including direct links to key documents (see: http://www.fordham.edu/academics/office_of_research/research_centers__in/center_for_ethics_ed/hiv_prevention_resea/research_ethics_regu_79380.asp).

Institutional Review Boards (IRBs) have become required at sites anticipating or receiving federal funds as a means to educate researchers and to ensure that research follows federal policy with regards to the ethical treatment of participants (Department of Health and Human Services [DHHS], 2005). When IRBs function well, they work collaboratively with investigators to develop good consent practices and monitor participant safety. The American Psychological Association [APA]'s ethical standards demand that when research requires approval by an institution, the information in the protocol must be accurate (APA, 2010, Standard 8.01). Paradoxically, however, at some institutions, overburdened, bureaucratic, or unresponsive IRBs may actually encourage deceit and create problems rather than expediting their resolution (Keith-Spiegel & Koocher, 2005). Unfortunately, charges that IRBs behave in unreasonable, unresponsive, and incompetent ways are common (Cohen, 1999; Giles 2005; Keith-Spiegel & Koocher, 2005). Still, investigators must carefully follow all IRB procedures at their institutions and obtain approval before enrolling participants.

FERPA, the Family Educational Rights and Privacy Act of 1974, establishes privacy rights of students at educational institutions receiving federal funds. Researchers seeking access to educational records will need to seek permission of parents (or students over age 18) to access such records with identifying data. Regulatory enforcement of FERPA occurs at the institutional level (i.e., a public school or university) rather than against specific individuals at such institutions.

HIPAA, the Health Insurance Portability and Accountability Act of 1996, has a significant bearing on the creation, storage, and use of protected health information (PHI). Federal regulations define PHI as oral, written, typed, or electronic individually identifiable information related to (a) a person's past, present, or future physical or mental health; (b) provision of health care to the person; or (c) past, present, or future payment for health care. Unlike FERPA, penalties for HIPPA violations target both the institutions and individuals responsible.

HIPAA also provides specific definitions of research and treatment (see 45 C.F.R. §164.501). The HIPAA definition of research describes "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." Treatment is described under HIPAA as "the provision, coordination, or management of health care and related services by one or more health care providers, including …. consultation between health care providers relating to a patient; or the referral of a patient for health care from one health care provider to another." HIPAA also distinguishes among subcategories of PHI, calling out the need for specialized authorization to release mental health records and psychotherapy notes. HIPAA also permits authorization for the use or disclosure of PHI in a single combined consent form along with other types of written information and permission for the same research. All mental health practitioners in the United States will become familiar with

HIPAA rules as part of their training or employment at health care settings; however, researchers must remain particularly sensitive when they seek, generate, store, or release research data that might qualify as PHI. In addition, research assistants and support staff having access to PHI will also require HIPAA training at the outset of their involvement in the project.

DSMBs, Data Safety Monitoring Boards, may be required under federal or institutional policy in certain circumstances. These groups consist of formally appointed independent (of the institution conducting the research) members assigned to conduct interim monitoring of accumulating data from research activities as a means to ensure the continuing safety of participants, the relevance of the study question, the appropriateness of the study, and the integrity of the accumulating data. Membership typically includes expertise in the relevant field of study, statistics, and research design. From the standpoint of clinical psychologists, DSMBs will most likely come into play in studies involving medical or pharmaceutical interventions, large multisite studies, and treatment protocols in which one arm of a study has potential inferiority or a need to halt the study early might otherwise occur. The DSMB will routinely monitor the progress of the study, protocol changes, and adverse events from a perspective independent of the investigators and their institutions.

Confidentiality Planning

One of the key issues to consider in advance of data collection involves planning how to protect the confidentiality of the participants in the study and how to share these plans with them. IRBs will expect investigators to develop and present a plan for this as part of the approval process and to describe the plan simply and clearly in the consent documents.

Certificates of confidentiality can be very important in cases of sensitive psychological research. Such certificates are authorized under the Public Health Service Act (1944–2006) and allow investigators to apply to the NIH and other DHHS agencies for such a certificate. Such certificates shield personally identifiable research records from forced disclosure by law enforcement or subpoena. As such, the certificates can help protect research information that might place participants in legal jeopardy or damage their financial standing, employability, insurability, or reputation (Fisher & Goodman, 2009)—for example, research in socially sensitive topic arenas, such as substance abuse and prostitution, or with populations that have faced social stigma, such as transvestites or persons with sexually transmitted diseases. The certificate will not excuse mandated reporters from their legal obligation to release confidential information as required by law in most states to protect children or other dependent persons from suspected abuse or neglect. Obtaining such a certificate does not prevent a researcher from voluntarily disclosing confidential information necessary to protect the research participant or others from harm (see APA, 2010, Standard 4.05, Disclosures).

Data storage and disposal issues should also be considered before beginning the project. We are required to protect the confidentiality of scientific information in all phases of record creation, maintenance, dissemination, and disposal (see APA, 2010, Standard 6.02). This would include data kept in the form of written and printed materials, automated scoring reports, audio and video recordings, websites or emails, company computer networks, storage on hard drives or disks, and facsimiles. Planning steps might include setting up a secured place to store the materials, limiting access to those members of the research team with a legitimate need to access them, coding the records or otherwise removing identifying data, and disposing of recordings or other records when they are no longer needed in accordance with any applicable legal requirements. Although some of these steps will not be carried out until the study is under way or has ended, information on such matters should become a part of any IRB filings and consent forms.

Consent to Participate

The concept of consent pervades our work as clinical psychologists. We seek consent to conduct assessments, consent for treatment, and consent for participation in research. The requirement of consent as a protection for participants first appeared in the Nuremburg code, published in the United States in the *Journal of the American Medical Association* in 1946. Adopted at the war trials of the 23 Nazi physicians indicted for crimes against humanity, the code includes the required elements of consent: legal capacity to give it, freedom to decline participation without coercion, and sufficient knowledge about the nature of the study to make an informed decision (Capron, 1989). Although never used as a legal precedent, the Nuremberg code forms the basis from which subsequent codes and policies were developed, including Standards 8.02 and 8.03

of the *APA Ethical Principles of Psychologists and Code of Conduct* (APA, 2010).

A brief detour into definitions seems warranted at this point. Many writers and regulations use the term *informed consent*, leaving open the question of how it might differ from *uninformed consent*. As used in the context of research participation, individuals may give consent only for themselves (and individuals for whom they serve as legally authorized guardians) and only when they do so competently, voluntarily, and with all the knowledge necessary to allow them fully informed choice. As described later, children or others who may not qualify under these terms may be asked to give their *assent* after a responsible parent or guardian has given his or her *permission*.

It is now well accepted that, with a few carefully proscribed exceptions, research participants must know what they are volunteering for and agree freely to do it. The consent issues that remain open or controversial in some respects focus on specific applications, such as how to deal with consent issues when conducting research online, when using deception paradigms, or when people may not have the ability to competently consent (e.g., suicidal or delusional individuals) or may not be able to provide consent for themselves (e.g., children) (Flewitt, 2005; Mishara & Weisstub, 2005; Rees & Hardy, 2003; Varnhagen, Gushta, Daniles, et al., 2005).

Voluntariness becomes a particularly sensitive issue because a person's decision to participate in a research project can be manipulated in both subtle and blatant ways based on intentional and unconscious factors. For example, solicitation by a person or institution perceived as prestigious and an authority can prove persuasive. Social influence becomes especially powerful if the researcher seems enthusiastic and likeable, or if the potential participant feels vulnerable, deferent, or in need of attention. People desperate for a solution to a personal matter related to the subject under investigation may overlook any risks. Some categories of potential participants such as inmates, students, or employees of the organization sponsoring the research may feel pressured to participate for fear of retribution, even when explicitly told that they are free to decline without penalty. Many of the people sought after for participation in the social and behavioral sciences are troubled, in need, or in a weaker bargaining position compared to the researchers.

This same well-recognized interpersonal dynamic applies to the right of participants to withdraw from a study after enrollment, even when they have no particular vulnerabilities (Milgram, 1963). Despite the disappointment that researchers undoubtedly experience when participants change their minds midcourse (especially if this occurs well into a complicated or longitudinal study), the right to withdraw should be honored (APA, 2010, Section 8.02). This right to disengage from a study should be made explicit during the initial consent phase. Rare exceptions may involve necessary interventions available only in a clinical research context (e.g., the need to taper off a medication rather than cease taking it completely at once).

The explicit offer of rewards, monetary or otherwise, can also have a potentially coercive influence on decisions about participating in research (Fisher, 2003; Oransky, Fisher, Mahadevan, & Singer, in press). Offering to pay participants a small amount to offset inconvenience and transportation costs will not likely be considered. Ethical issues arise when the reimbursements or rewards for participating are significant enough to influence consent decisions (see APA, 2010, Standards 8.06a and 8.06b).

Some participants may discount any potential risks in the hope of securing needed benefit, or they may believe that needed services are contingent on participation in research. Researchers must be careful not to engage in "hyper-claiming"—that is, suggesting to potential participants that the study will reach goals that are, in fact, unlikely to be achieved (Rosenthal, 1994).

Researchers should never create guilt feelings in those who decline to participate in a research project. In its nastiest form, a researcher may actually hint that refusal to participate suggests selfishness or lack of caring about others who could benefit from the "good people" who have agreed to be part of the study. Another form of such subtle coercion involves appealing to the participant's altruism. Such requests can range from personal pleas for help to suggestions that cooperation will benefit humankind or advance science. To the extent that researchers genuinely need participants and are sincere in their beliefs that they are doing worthy work, some level of altruistic appeal is probably unavoidable. However, as the example presented in the next section illustrates, some potential participants are more gullible than others when appeals to altruism are employed.

Legal capacity comes into play when a potential participant lacks either the legal standing (e.g., minors) or the psychological competence (e.g., people with developmental disabilities, dementia, or severe mental illness) to grant consent. In such

instances we turn to proxy consent strategies by seeking permission from legally authorized guardians and assent by the potential participant, if appropriate to do so.

As an illustration of children's potential willingness to agree to outrageous requests in the context of research, Keith-Spiegel and Maas (1981) appealed to the altruism of school-aged children, asking whether they would participate in highly unusual research. When told that participating in an experiment would help starving children from dying by the thousands, 80 percent of school-aged children agreed to eat a bite of baked mouse. No baked mouse was ever served to the children, of course, because the point of the study was to evaluate altruistic appeals. But the surprising level of effectiveness of a fervent altruistic appeal on a vulnerable population was clearly demonstrated. In the same study, a majority of the children would also agree to have their eye poked with a glass rod when the research was described as helping blind children see again.

After securing permission for the legally incompetent individual from a parent or guardian, federal regulations (DHHS, 2005, 45 C.F.R. § 46.408a) and the APA ethics code (APA, 2010, Section 3.10b) encourage a respectful quest for the person's assent to participate using language appropriate to his or her developmental level. However, in some instances assent may not be appropriate (e.g., if the potential participant lacks all communication abilities or if participation holds a genuine prospect of direct benefit to the person's health and is available only in the context of the research).

In some circumstances minors may agree to participate in research without the permission of a parent or guardian, as described in 45 C.F.R. § 46.402a (DHHS, 2005). Such circumstances may involve so-called "emancipate" or "mature minors," who have taken on adult responsibilities such as marriage or parenthood. In some states where minors may consent to treatment for sexually transmitted diseases, substance abuse, or similar conditions without parental consent, IRBs may waive parental permission requirements. Unfortunately, IRBs often seem reluctant to exercise this option, resulting in the inability to study some life-threatening social conditions of adolescence (Fisher & Goodman, 2009). IRBs may also waive the permission requirement when the research involves no more than minimal risk, although—as will be discussed later in this chapter—the definition of minimal risk is not always clear.

KNOWLEDGE AND UNDERSTANDING

For research participants to understand what they are being asked to agree to do, they (or their surrogates) must have the capacity to comprehend and evaluate the information offered to them prior to enrolling. Gaining fully informed consent is best viewed as a critical communication process during which an agreement is reached. But many people may not have sufficient self-awareness or candor to admit that they do not understand something, especially if they do not feel in control of the situation. Researchers must understand the process of obtaining consent and understand that a signed form is *not* synonymous with informed consent. If people do not fully understand what they have signed, true consent has not occurred. Unfortunately, a number of studies have documented that many legally competent adults have minimal understanding of what they agree to do as participants in research (e.g., Cassileth, 1980; Sieber & Levine, 2004; Taub, Baker, & Sturr, 1986).

Individuals who have trouble with the language of the person seeking consent, for whatever reason, require special consideration. Non-English-speaking participants have the right to appropriate translation and interpretation. Individuals with poor reading skills should also receive special assistance. All participants should be made to feel comfortable asking questions to preclude attempts to "save face" by agreeing to participate in an activity that is not fully understood or to please those with perceived higher social status and authority.

When participants lack legal capacity to give consent, permission (i.e., proxy consent) must be obtained from authorized others. Nevertheless, except for infants, nonverbal children, and the seriously impaired, participants should be offered some explanation of what they are being asked to do. In addition, if practical, their proactive assent should be sought. It is our position that even if permission has been obtained, an individual who expresses lack of desire or interest in participating should be excused unless there is a very compelling reason, such as a likelihood of direct therapeutic benefit, to override the participant's wishes (Koocher & Keith-Spiegel, 1990, 2008).

Consent forms also protect researchers and their institutions, allowing for a "record of agreement" should the participant complain later (Sieber & Levine, 2004). Also, formal informed consent procedures are not always required for some types of data-collection methods, such as so-called "minimal risk" research when the project is highly unlikely to cause any harm or distress. Examples include anonymous

questionnaires, naturalistic observations, and some types of archival research or review of data collected for nonresearch purposes with participant identities removed (APA 2010 Standard 8.05; DHHS, 2005). In addition, formal consent agreements are typically unnecessary for service and program evaluations in educational settings or for job or organizational effectiveness so long as there is no risk to employability and confidentiality is protected.

Goodness-of-Fit Ethics

Because so many different variables related to a particular study, context, or participant sample may interact to complicate ethical decision making in research, Fisher and her colleagues (Fisher, 2002; Fisher & Goodman, 2009; Fisher & Vacanti-Shova, in press) have described a highly creative goodness-of-fit model. They conceptualize research vulnerability (susceptibility of participants to research risk) as an interaction of the sample population characteristics, individual participant characteristics, and the research context. Their model strives to minimize harm by fitting the research procedures to the participants' susceptibilities. As an example, Fisher and Vacanti-Shova (in press) cite the misconceptions and mistrust associated with HIV vaccine clinical trials among intravenous drug users. She describes development of population-sensitive educational materials that enabled a process she calls "co-learning" to occur between investigators and potential participants, leading to an increase in knowledge and trust.

Addressing the Therapeutic Misconception

Appelbaum and his colleagues (Appelbaum, Lidz, & Grisso, 2004; Appelbaum, Roth, & Lidz, 1982) identified and elaborated on the important phenomenon of a therapeutic misconception to describe an all-too-common but incorrect belief by people participating in randomized clinical trials (RCTs). Even after receiving careful consent information explaining random assignment, many participants still assume that somehow their individual needs will come into play as part of their assignment to treatment, control, or comparison groups. Many also cling to an unreasonable expectation of medical benefit from their research participation.

This phenomenon (i.e., the assumption that something therapeutic will flow from one's participation in an RCT) may represent a kind of blind optimism or hope combined with the positive social valence assigned to health care or scientific researchers by many people in society. The invitations to participate often come from investigators based at universities or teaching hospitals. The recruiting personnel often develop attentive personal communication relationships with potential enrollees as they provide study-related information and explanations. In addition, the candidates for study often have treatment-resistant or refractory conditions. As a result, the expectations or beliefs of potential participants may go well beyond the facts conveyed.

APA's Ethical Standard 8.02b, on informed consent to research, requires psychologists to address such potential misconceptions during the consent process. Key to doing so involves explaining that "experimental" treatment does not necessarily mean better treatment or treatment with known direct benefits for participants. This may require showing special attentiveness and providing clarifying information to patients who show signs of misunderstanding.

Clinical Equipoise

When setting up RCTs, the fundamental nature of the research often involves exploration of efficacy and effectiveness in ways that anticipate some participants will have a worse outcome than others. To reject the null hypothesis, we most likely expect that those receiving a particular treatment or research condition will fare better than those in the alternative conditions. We should therefore ethically require an expectation of clinical equipoise (i.e., genuine uncertainty regarding the comparative merits of the different treatment arms to which participants may find themselves randomized).

We should stand prepared to ensure that control group participants will ultimately have access to an available treatment that proves effective and that people participating in the experimental treatment arm of the study will not face a greater risk than they would had they been assigned to a standard care paradigm or the control group (Fisher & Vacanti-Shova, in press). In mental health research this ethical stance requires us to consider critical periods in the trajectory of human development. For example, we must ask ourselves whether nonintervention during a critical clinical or developmental interval will lead to permanent changes or disadvantage that cannot be remediated by offering those in the control group the intervention that proved superior at a later point in time.

Compensation of Participants

Some research protocols seek to enhance enrollment or retain participants longitudinally by offering compensation (e.g., cash, prizes, services, or

eligibility for these). The consent process must accurately inform potential enrollees under what conditions they will qualify for all, some, or none of the potential compensation. In addition, APA Ethical Principle D (Justice) and Standard 8.06 require us to ensure that any inducements to participation that we offer do not rise to the level of economic coercion. This can become particularly relevant for economically disadvantaged populations or those without access to mental health services, when an RCT focused on such treatment is involved (Fisher & Vacanti-Shova, in press; Koocher, 2005).

Children and Adolescents as Special Cases

As noted earlier, children will often lack the legal status and cognitive or emotional maturity necessary to provide fully competent consent—leading us to rely on guardian permission and child assent paradigms. Raising this concern once again underscores a particular set of issues that come up when compensation may be offered to parents or guardians in exchange for permitting the participation of their children. The assent of the child (i.e., the right of the child to veto participation or withdraw from research participation that parents or guardians might permit) must be weighed carefully with any potential risks or direct benefits to the child (Kendall & Suveg, 2008). If a study holds no promise of direct benefit, the child should have an absolute right to refuse participation.

In one egregious example, families were asked to enroll the younger siblings of incarcerated juvenile offenders in a drug study aimed at identifying children at risk for antisocial behavior. That study inappropriately played on parental guilt regarding the incarcerated sibling, offered significant financial inducement to economically disadvantaged families, and placed participating children at some risk of acquiring stigmatizing labels, in addition to exposure to the drug fenfluramine (Koocher, 2005).

The key to good practice involves focusing special attention on the vulnerabilities and preferences of those who might not otherwise qualify to exercise consent for themselves. Investigators should consider such matters in the design of their research, and IRBs overseeing the approval of such projects should have or bring in the expertise necessary to address such concerns.

Anonymous, Naturalistic, or Archival Research

Many opportunities arise to conduct psychological research on populations for which "disclosure of responses would not place participants at risk" (APA, 2010, Standard 8.05). Such research might include the use of anonymous questionnaires, naturalistic observations, or delving into archival records and databases. In situations where confidentiality is protected, when disclosure would not place participants at risk (i.e., using the CABLES categories), and when the research cannot reasonably be expected to cause harm (including emotional distress), IRBs may waive the requirement for informed consent. At the same time, careful protocol criteria should specify ways in which data collected from identifiable archives (i.e., old patient records) will be sanitized and protected.

When considering Internet-based data collection, including both individualized survey invitations and observational research (e.g., monitoring chat rooms, Twitter feeds, or Facebook postings), investigators should pay close attention to the requirements of APA's (2010) Standards 3.10 (consent) and 8.02 (research consent). Participants solicited individually can obviously choose not to respond, after reading the applicable consent information. However, special attention to avoid capturing or storing identifying data is required for observational studies based on data acquired surreptitiously (e.g., collected by lurking and recording traffic on public networking sites).

Conflicts of Interest and Research Funding

Although medicine has suffered far more than psychology from scandals associated with industry-sponsored research leading to conflicts of interest, the mental health professions have included some of the most dramatic cases. As an example, a psychiatrist who will remain unnamed here has become a poster child for bad behavior mixing industry and scientific interests in mental health research. Using the Google search engine and the terms "[person's name] conflict of interest" yielded more than 23,000 hits in July 2010. Readers will draw their own conclusions, but the message has not been lost on the keepers of scientific integrity. Academic institutions and scientific publications increasingly demand full transparency of all potential conflicting interests in the conduct and publication of research.

When accepting any financial or in-kind support for research, investigators must exercise full transparency and consider the multiple role conflict that occurs "if the multiple relationships could reasonably be expected to impair the psychologist's objectivity, competence, or effectiveness in performing his or her functions as a psychologist" (APA, 2010,

Section 3.05a). Full disclosure of any such support would be expected at any public presentation of the findings in oral or written format.

Midcourse Corrections

To this point, the chapter has focused on ethical issues one can conceptualize and incorporate in planning prior to initiating data collection. Once the research gets under way, many new ethical challenges arise and may dictate changes in the *modus operandi* of the project.

Data Collection and Recordkeeping

As noted earlier, most researchers will rely on assistants, technicians, or students at the front line of interaction with participants. This may involve reviewing consent forms with participants, obtaining signatures, conducting or recording interviews, administering tests, scoring protocols, filing, computerized data entry, and other such operational tasks. The supervising psychologist retains ethical responsibility for training these assistants and overseeing their work (see APA, 2010, Standard 2.05). Note that at times the responsible psychologist will be the faculty member who agrees to act as the chair of a student's thesis or dissertation.

Oversight of assistants or supervisees includes responsibility for training the assistants as well as their routine work on the project. Training of assistants should include discussion of data confidentiality requirements and protocols for issues that might trigger the need to contact their supervisor urgently (e.g., if a research participant discloses information that may require a mandated breach of confidentiality or threatens harm to himself or herself or another). Assistants must also remain mindful of participants' right to withdraw, and should be prepared to address or call for help with other unexpected events that occur (e.g., the child who vomits on test materials or the adult participant who arrives intoxicated).

Data Safety Monitoring

Formal DSMBs may be required in some studies. However, unexpected adverse events can come up at any time, even in studies where one could not have reasonably anticipated them. For example, an undergraduate research project involved asking elementary school children to draw pictures of "What you want to be when you grow up." One 7-year-old girl drew a police officer and told the stunned student researcher that she'd arrest people who abused their children as her father abused her. Those who discover or witness such events should bring them to the attention of the responsible researcher immediately. In other long-running studies, reports of adverse incidents or interim analyses have revealed one treatment group faring much better than another. In such instances, interruption of the study for the benefit of all participants may prove warranted.

Scientific Misconduct

The scientific enterprise is built on the premise that truth seeking is every researcher's principal motivation. Those who review submitted grants or papers for publication accept on faith that researchers subscribe to the highest standards of integrity (Grinnell, 1992). And, fortunately, researchers themselves value the ethical standards to which they are held (Roberts & McAuliffe, 2006). Unfortunately, other motives have prompted some researchers to cheat.

Recurring themes in unmasked data fraud involve perpetrators who were lax in the supervision of the data gathering and analysis and excessively ambitious with previous records of prolific writing. Usually present is an intense pressure to produce new findings. Researchers who publish first are credited with a "discovery," and showing good progress is essential to continuing grant funding. In fact, one reason many researchers may jump ahead of their actual findings by reporting bogus or manipulated data is because they sincerely believe that they already know what the factual findings will be. Their commitment to a theory or hypothesis may be so strong that it even diminishes fears of detection (Bridgstock, 1982). These perpetrators of scientific fraud believe in their hearts that doing wrong now will prove right later.

Unfortunately, additional elements present in the scientific enterprise can tempt researchers to cheat. A project that fails to produce statistically significant findings may not gain acceptance for publication due to a bias against publishing statistically insignificant findings. Or, legitimate data may never get reported because a financial interest in a particular outcome failed to materialize. The organizational culture in which researchers conduct their studies can also contaminate science. That is, when a researcher sees others behaving unethically as a way of getting ahead in a competitive environment, and those with the authority to take action turn a blind eye, undue moral pressure is exerted on those who would otherwise behave ethically (Keith-Spiegel, Koocher, & Tabachnick, 2006). Sometimes

those in authority behave in an overly restrictive, unresponsive, biased, unfair, or offensive manner that inhibits the ability to do sound research, thus inviting rule breaking.

TYPES OF MISCONDUCT

The two most serious and often discussed forms of scientific misconduct are *fabrication* and *falsification*. Fabrication usually takes the form of "dry lab" data that are simply invented. Falsification can take several forms. Actual data can be "smoothed" or "cooked" to approach more closely the desired or expected outcome. Or collected data points can be dropped or "trimmed" to delete unwanted information.

Similar to plagiarism, the purposeful creation of unsound data is considered among scientists and scholars as a grievous ethical violation (see APA, 2010, Standard 8.10a). The consequences of making invalid data public is, however, far more serious than simply duplicating the work of others (i.e., plagiarism) because the spurious conclusions contaminate the research record. Conducting good science requires a process of building on previous work. Time and effort becomes wasted when trusting researchers pursue inquiries based on previously reported findings they do not realize are bogus. Application of findings based on tainted data can even cause harm. For example, if a researcher proposing an experimental therapy technique "trims" data and the tainted findings appear in a reputable journal, the results may be applied by unsuspecting mental health professionals. By the time someone notices that clients are not improving (or their condition is worsening), serious setbacks could occur. Or if a developer of a psychodiagnostic assessment "cooks" the validity data, once published and in use, people could be misclassified using what is believed to be a fair test. To the extent that such test results are used to determine diagnoses or treatment approaches, or to determine who should not be hired, serious errors are committed that have a deleterious impact on people's lives.

Several much less frequently discussed questionable acts can also distort the scientific record to the same degree as fabrication or falsification. It is possible to set up an experimental condition so that the collected data are more likely to confirm a hypothesis. For example, a researcher may purposely select those with only the *mildest* symptoms of a diagnostic category to bolster the chances of "proving" that a particular therapy works. Or, conversely, participants with the *most severe* symptoms may be purposely placed in the control group so that the experimental group will appear to fare more favorably by comparison. Biased reporting of results, such as presenting the findings in such a way that they appear far more significant than they actually are, misleads readers. Relying on secrecy to get ahead, refusing to share data, and withholding details in methodology or results run against the grain of scientific integrity because they make it more difficult or impossible for anyone else to successfully pursue that same line of inquiry (Grinnell, 1992; Martinson, Anderson, & de Vries, 2005; Sieber, 1991a, 1991b). "Data torturing" is another ignoble practice, involving analyzing the same data many different ways until one finds statistical significance (Whitley, 1995). These ethically questionable acts may be more likely to occur when the source of financial support has an interest in obtaining findings favorable to its desired outcome.

Sometimes invalid data were not purposely created; ineptitude may be the issue. Incompetence can result in inappropriate design, poor or biased sampling procedures, misused or wrongly applied statistical tests, inadequate recordkeeping, and just plain carelessness. Even though there may be no intent to deceive, inaccurate information can also seriously damage the research record.

One may be tempted to assume that such inaccuracies, purposeful or not, will be discovered. But we cannot count on it. Whereas errors in alleged scientific advances are assumed to be eventually self-correcting through replication, funding sources typically do not support replication research. Furthermore, most scholarly journals do not normally publish replication studies. Thus, there is little incentive for researchers to repeat studies, especially those that were expensive and complex.

INCIDENCE

We want to believe that most professionals who conduct research are overwhelmingly honest. However, in a now-classic survey of doctoral candidates and faculty members from 99 departments, anonymous responses to questions about knowledge of instances of scientific misconduct revealed that over two thirds of the graduate students and about half of the faculty had direct knowledge of the commission of some form of scientific misconduct (several people may be aware of one event and many people may have essentially reported on the same event). Most did not confront or report it, usually from fear of reprisal (Swazey, Anderson, & Lewis, 1993). A more recent study suggests that the situation has not improved. Although very few of the

several thousand scientists responding to an anonymous survey disclosed committing more serious research sins of fabrication or falsification, one out of every three admitted to committing an act that could be labeled as questionable (Martinson, Anderson, & de Vries, 2005). Thus, unfortunately, poor conduct may be neither rare nor ever adequately resolved, and not all bad behavior equates to guilt.

DIFFICULTIES IN DETECTION

Most of the highly publicized data scandals occur in biomedical research laboratories. No one knows for sure whether the incidence is higher in biomedical science than in social and behavioral science, or whether it is simply easier to detect fraud in biomedicine. Most social and behavioral research does not involve chemical analyses, tissue cultures, change in physical symptoms, invasive procedures, or similar "hard" documentation. Social science data, on the other hand, often take the form of numerical scores from questionnaires, psychological assessments, performance measures, or qualitative data based on interviews or behavioral observations. The actual research participants have long since gone, taking their identities with them. Such data are relatively easy to generate, fudge, or trim. We can hope that social science researchers are motivated by the responsible quest for truth, but it is disquieting to note that the same publish-or-perish and grant-seeking pressures exist for social and behavioral scientists working in competitive settings, that fame is an ever-present allure in any field, and that the practice of fabricating data may start when researchers were students (e.g., Kimmel, 1996; Whitley & Keith-Spiegel, 2002).

Ethics committee experiences are not very helpful in estimating the incidence of publishing fraudulent data in the social and behavioral sciences. The rare charges of falsifying data brought to the attention of ethics committees have proved difficult to adjudicate, as the next cases reveal.

Does actual harm result only from biomedical research fraud? Not necessarily. Dishonest social and behavioral scientists can seriously disadvantage people as well. In a case that came to light in the late 1980s, the federally funded research of psychologist Stephen Breuning reported findings based on data that were never collected. This case is especially disturbing because Breuning's findings became a basis for treatment decisions before they were discredited. The fraudulent reports were also then used as a basis to determine drug therapy for institutionalized severely retarded persons, a treatment later proven detrimental based on the results of competent research conducted by others (Bell, 1992; Committee on Government Operations, 1990).

In light of such disturbing reports, what credibility should the public place on researchers' work? The public's generalized distrust, which is sure to ensue as the media prominently exposes cases of scientific intrigue, could have disastrous consequences for everyone. Scientists are as dependent on the public for continued support as society is on their valuable, legitimate contributions. Although part of the problem is a system of rewards that implicitly encourages dishonest and irresponsible scientific practices, researchers must remain true to the search for truth if the entire scientific enterprise is to remain stable and healthy. In response to a growing concern about scientific dishonesty, the Public Health Service established the Office of Research Integrity (ORI) to direct activities focused on research integrity on behalf of DHHS. This office creates policies and regulations with regard to preventing, detecting, and investigating scientific misconduct (see http://ori.dhhs.gov/_).

After-Effects

Once data collection has ended and participants have exited the study, the nature of ethical issues typically encountered shifts significantly. As described earlier, the investigators do owe participants continued duties of confidentiality, as well as obligations to secure, maintain, and properly dispose of identifiable materials. Scientific misconduct (e.g., fraud, fabrication, and plagiarism) that often begins during earlier stages of the research enterprise may also continue.

Scholarly Publishing Disputes

Knowledge is shared and advanced through scholarly books and journals and, with increasing frequency, over the Internet. The primary purpose of scholarly publishing outlets is to disseminate useful discoveries as soon as practicable, sometimes as quickly as a few months following the completion of the study. Despite what one might assume functions as a sophisticated and collaborative process, scientific writing and research publication is, in fact, fraught with the potential for intense conflict and spiteful disputes. Why do smart people have such problems in this arena? The main reason is that the stakes are very high for those who want to advance their careers. Publication credits are often required to gain entrance into graduate school or to land an attractive postdoctoral appointment, to obtain or

retain a job, to earn a promotion, or to win grant funding. Publications also elevate researchers' status among their peers. And, whereas publications used to carry no direct monetary gain, scientific findings can now become the basis for profit-making partnerships with business ventures.

The competition to "get published" can interject unhealthy features into the scientific enterprise. A focus on quantity rather than quality may prompt some researchers to pursue projects that can be completed rapidly rather than tackling more noteworthy undertakings or studying a subject matter in more depth.

Publication credit may seem like a relatively straightforward procedure in terms of deciding who deserves authorship credit and in what order to list multiple contributors. Not so. Research has become more specialized in recent years, often requiring teams composed of many people who have no or minimal overlapping skills. Although quite rare, a single article may list more than 100 authors (McDonald, 1995). Because of the potential boost to one's career, bitter disputes over the assignment of publication credits have occurred, and differences of option have become increasingly common. Over a quarter of the respondents to a large survey believed that they had fallen victim to unfair or unethical authorship assignments (Sandler & Russell, 2005).

Senior (first-listed) authorship is the most coveted position. But why the fuss over whose name appears first? The first-listed individual is presumed to be the major contributor and the name by which the work will be indexed (Fine & Kurdek, 1993). "Junior" (second and later-listed) authors have become upset when individuals—usually those with the power and authority over them—claim the senior authorship for themselves, even though they had only minimal involvement in the project (Holaday & Yost, 1995). Some contributors complain that they received mere mention in a footnote or no acknowledgment at all when their involvement actually warranted a junior authorship.

Ethics committees have agreed that sometimes more powerful or exploitative researchers disadvantaged junior-listed authors. But at other times honest differences in opinion about the value placed on each others' contributions are at issue. In some instances graduate students have alleged that thesis and dissertation supervisors insisted on credit as coauthors on any published version of the students' projects. Many students appear to view their supervisors as fulfilling the obligation to facilitate their professional development, while supervising professors may see their contributions as essential to ensuring acceptable publication quality. At the same time, almost every senior researcher has at least one story to tell about a student who abandoned what started out happily as a joint research venture. Ideally, research collaborators should reach agreements about what each person can reasonably expect in terms of credit before the research or writing collaboration begins, and addressing any need for modifications should changes in the project's status occur (Hopko, Hopko, & Morris, 1999).

Ethical problems in publication credit also occur in the opposite direction. "Gift authorships," offered as a favor to enhance a student's application to graduate school or to advantage a nontenured colleague, appear to be a generous gesture. However, to the extent that the authorship was unearned, others may be unfairly disadvantaged. For example, in a competitive employment situation, the applicant with an unearned authorship could unfairly prevail over others who were just as (or more) qualified. Assigning authorship credit to a well-known senior researcher who had minimal involvement in the work for the purpose of possibly enhancing the potential for publication constitutes another form of ethically unacceptable gift authorship. Many journal editors now require anyone submitting a manuscript to specify the roles of each named co-author in the project. For example, the online instructions to authors for the prestigious journal *Nature* note: "Authors are required to include a statement of responsibility in the manuscript that specifies the contribution of every author."

Professional ethics codes do address these issues. The APA ethics code (APA, 2010, Standards 8.12a and 8.02b) specifies that authorship credits are to be assigned in proportion to authors' actual contributions. Minor or routine professional contributions or extensive nonprofessional assistance (e.g., typing a complicated manuscript, coding data, or helpful ideas offered by a colleague) may be acknowledged in a footnote or in an introductory statement (APA, 2010, Standard 8.02b). Publications arising from students' theses or doctoral dissertations should normally list the student as the senior author, even when advisors or others were heavily involved in the project (APA, 2010, Standard 8.12c).

Data Sharing

Computers and electronic transfer systems allow inexpensive data banking and instant data sharing anywhere in the world in a manner never envisioned a couple of decades ago. Investigators have

an ethical obligation to preserve and share their research data with other investigators under reasonable circumstances. Data sharing among scientists holds the potential for hastening the evolution of a line of inquiry, helps to ensure validity and error corrections, encourages collaborative ventures, and is generally encouraged when done responsibly (APA, 2010, Standard 8.14). Federal regulations specify some parameters for such sharing when the research has taken place with government support (see http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm). Many scientific journals also provide for such sharing by requiring authors to maintain the data on which published articles are based for a period of years.

Even research participants receive an advantage in the sense that their contributions are maximized. However, concerns about privacy invasion have drastically increased as technological advances allow sophisticated surveillance as well as links and access among computer storage banks. Researchers should ideally work with their IRBs to store and encode data in a manner that protects participant privacy from the outset, and resist opportunities to contribute information to data banks if participants' confidentiality cannot be safeguarded.

Researchers also have legitimate interests in benefiting from their scientific work and may reasonably withhold sharing data pending the publication of results. When large epidemiologic or longitudinal studies collect data over several discrete time periods or waves, it is reasonable to expect release of data in waves as main findings from each wave are published.

Participant Debriefing

To the extent that research participants have an ongoing interest in the results of a study, they should be afforded an opportunity to obtain such information (Jeffery, Snaith, & Voss, 2005). In many cases participants may have no such interest, while in other cases the clinical findings may offer significant benefit to some. These matters deserve full consideration at the time a project is reviewed by the IRB, and the consent process should advise participants of their options in this regard. To the extent that investigators have promised such follow-up, they have an ethical obligation to follow through and deliver on that promise.

Conclusion

The ethical execution of clinical research requires careful thought and planning at each step of the enterprise. This includes consideration of risks and benefits to participants and full collaboration with others charged with regulating the research enterprise. Participants should enter the project with a full and clear understanding of what is expected of them and what they can expect from the research team in return. As the work moves forward, the well-being of participants demands constant monitoring in full conformity with applicable law and prior assurances. When the research data are collected, investigators must follow through on all promises made to the participants, including the protection of their privacy. Investigators must also adhere to appropriate standards of scientific integrity in the analysis of their data, presentation of their findings, and data sharing following publication.

Future Directions

Many interesting ethical questions remain for future exploration. One important subset deals with technological innovation and the ability to use new techniques in monitoring participants and collecting data. We will assess, treat, and study people via remote telemetry that opens wondrous new sampling opportunities, while simultaneously raising new privacy and integrity concerns. Can we ensure the privacy of people who provide personal information via the Internet or other forms of distant communication? Can we feel confident of the identity of the person who purports to provide the data? Can we fulfill our obligation to intervene with research participants deemed at risk when the investigators are not in physical proximity to the participants?

Technology has already created significant ethical problems related to fraud in the biomedical community. For example, using photographic enhancement software has led to unethical manipulation of tissue slides and similar illustrations in experimental biology and medical journals, requiring new publication standards for accepting such material. Just as one can now easily fabricate or falsify such images, we will need to guard against similar technologically driven dishonesty in the behavioral sciences. Plagiarism has become easier thanks to Internet searches, but it has also become more readily detectable thanks to text comparison software. There will be new ways to cheat and new ways to detect cheating in the context of research.

Another important subset of ethical challenges will arise as more psychologists qualify for prescribing privileges and engage in pharmaceutical research. Such psychologists will face some of the

same ethical challenges as physicians, who have too often become tools of the pharmaceutical industry.

As we move forward, the details of how we conduct our research will continue to evolve, but the underlying ethical principles will remain unchanged: honesty, integrity, and respect for the people we work with will remain paramount.

References

American Psychological Association (2010). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.

Annas, G. J., & Grodin, M. A. (1995). *The Nazi doctors and the Nuremberg Code: Human rights in human experimentation*. New York: Oxford University Press.

Appelbaum, P. S., Lidz, C. W., & Grisso, T. (2004). Therapeutic misconception in clinical research: Frequency and risk factors. *IRB: Ethics and Human Research*, *26*(2), 1–8. doi:10.2307/3564231

Appelbaum, P. S., Roth, L. H., & Lidz, C. (1982). The therapeutic misconception: Informed consent in psychiatric research. *International Journal of Law and Psychiatry*, *5*, 319–329. doi:10.1016/0160–2527(82)90026–7

Bayer, R. (1989). *Private acts, social consequences: AIDS and the politics of public health*. New York: Free Press.

Bell, R. (1992). *Impure science*. New York: Wiley.

Boudewyns, P. A., & Shipley, R. H. (1983). *Flooding and implosive therapy: Direct therapeutic exposure in clinical practice*. New York: Plenum.

Bridgstock, M. (1982). A sociological approach to fraud in science. *Australian & New Zealand Journal of Statistics*, *18*, 364–383.

Burris, S., & Gostin, L. (1997). Genetic screening from a public health perspective: Some lessons from the HIV experience. In M. A. Rothstein (Ed.), *Privacy and confidentiality in the genetic era* (pp. 137–158). New Haven, CT: Yale University Press.

Capron, A. M. (1989). Human experimentation. In R. M. Veatch (Ed.), *Medical ethics* (pp. 125–172). Boston: Jones & Bartlett.

Cassileth, B. R. (1980). Informed consent—why are its goals imperfectly realized? *New England Journal of Medicine*, *302*, 896–900.

Cohen, J. (Nov., 1999). The federal perspective on IRBs. *APS Observer*, *5*, 19.

Committee on Government Operations. (1990). *Are scientific misconduct and conflicts of interest hazardous to our health?* (House Report 101–688). Washington, DC: Author.

Darou, W. G., Hum, A., & Kurtness, J. (1993). An investigation of the impact of psychosocial research on a native population. *Professional Psychology*, *24*, 325–329.

Department of Health and Human Services (2005). Title 45: Public Welfare, Part 46, *Code of Federal Regulations, Protection of Human Subjects.* Accessed June 30, 2011, at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

Family Educational Rights and Privacy Act of 1974 (FERPA), 20 U.S.C. § 1232g, Accessed July 1, 2011, at http://www.law.cornell.edu/uscode/20/1232g.html.

Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist*, *48*, 1141–1147.

Fisher, C. B. (2002). A goodness-of-fit ethic of informed consent. *Fordham Urban Law Journal*, *30*, 159–171.

Fisher, C. B. (2003). Adolescent and parent perspectives on ethical issues in youth drug use and suicide survey research. *Ethics & Behavior*, *13*, 303–332. doi:10.1207/S15327019EB1304_1

Fisher, C. B., & Goodman, S. J. (2009). Goodness-of-fit ethics for non-intervention research involving dangerous and illegal behaviors. In D. Buchanan, C. B. Fisher, & L. Gable (Eds.), *Research with high risk populations: balancing science, ethics, and law* (pp. 25–46). Washington, DC: American Psychological Association. doi:10.1037/11878–001

Fisher, C. B., & Vacanti-Shova, K. (2012). The responsible conduct of psychological research: An overview of ethical principles, APA Ethics Code standards, and regulations. In S. Knapp (Ed.), *Handbook of ethics in psychology* (pp. 333–368). Washington DC: American Psychological Association

Fisher, C. B., & Wallace, S. A. (2000). Through the community looking glass: Re-evaluating the ethical and policy implications of research on adolescent risk and psychopathology. *Ethics & Behavior*, *10*, 99–118.

Flewitt, R. (2005). Conducting research with young children: Some ethical considerations. *Early Child Development and Care*, *175*, 553–565.

Gamble, V. N. (1993). A legacy of distrust: African Americans and medical research. *American Journal of Preventive Medicine*, *9*, 35–38.

Giles, J. (2005). Researchers break the rules in frustration at review boards. *Nature*, *438*, 136–137.

Grinnell, F. (1992). *The scientific attitude* (2nd ed.). New York: Guilford Press.

Harris, Y., Gorelick, P. B., Samuels, P., & Bempong, I. (1996). Why African Americans may not be participating in clinical trials. *Journal of the National Medical Center*, *88*, 630–634.

Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No. 104–191, 110 Stat. 1936. Accessed July 1, 2011, at https://www.cms.gov/HIPAAGenInfo/Downloads/HIPAALaw.pdf

Holaday, M., & Yost, T. E. (1995). Authorship credit and ethical guidelines. *Counseling and Values*, *40*, 24–31.

Hopko, D. R., Hopko, S. D., & Morris, T. L. (1999). The application of behavioral contracting to authorship status. *Behavior Therapist*, *22*, 93–95.

Hudson, K. L., Rothenburg, K. H., Andrews, L. B., Kahn, M. J. E ., & Collins, F. S. (1995). Genetic discrimination and health insurance: An urgent need for reform. *Science*, *270*, 391–393.

Jeffery, A., Snaith, R., & Voss, L. (2005). Ethical dilemmas: Feeding back results to members of a longitudinal cohort study. *Journal of Medical Ethics*, *31*, 153.

Jones, J. (1981). *Bad blood*. New York: Free Press.

Keith-Spiegel, P., & Koocher, G. P. (2005). The IRB paradox: Could the protectors also encourage deceit? *Ethics & Behavior*, *14*, 339–349.

Keith-Spiegel, P. Koocher, G. P., & Tabachnick, B. (2006). What scientists want from their research ethics committees. *Journal of Empirical Research on Human Research Ethics*, *1*, 67–81.

Keith-Spiegel, P., & Maas, T. (1981). *Consent to research: Are there developmental differences?* Paper presented at the annual meetings of the American Psychological Association, Los Angeles.

Kendall, P. C., & Suveg, C. (2008). Treatment outcome studies with children: Principles of proper practice. *Ethics and Behavior, 18*, 215–233.

Kimmel, A. J. (1996). *Ethical issues in behavioral research*. Cambridge, MA: Blackwell Publishers.

Koocher, G. P. (2002). Using the CABLES model to assess and minimize risk in research: control group hazards. *Ethics & Behavior*, *12*, 75–86.

Koocher, G. P . (2005). Behavioral research with children: the fenfluramine challenge. In E. Kodesh (Ed.), *Learning from cases: ethics and research with children* (pp. 179–193). New York: Oxford University Press.

Koocher, G. P., & Keith-Spiegel, P. C. (1990). *Children, ethics, and the law: Professional issues and cases.* Lincoln, Nebraska: University of Nebraska Press.

Koocher, G. P., & Keith-Spiegel, P. C. (2008). *Ethics in psychology and the mental health professions: standards and cases (3rd ed.)*. New York: Oxford University Press.

Lapham, E. V., Kozma, C., & Weiss, J. O. (1996). Genetic discrimination: Perspectives of consumers. *Science*, *274*, 621–624.

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, *435*, 737–738.

McDonald, K. A. (April 28, 1995). Too many co-authors. *Chronicle of Higher Education*, A35.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Mishara, B. L., & Weisstub, D. N. (2005). Ethical and legal issues in suicide research. *International Journal of Law and Psychiatry*, *28*, 23–41.

Moreno, J., Caplan, A. L., & Wolpe, P. R. (1998). Updating protections for human subjects involved in research (policy perspectives). *Journal of the American Medical Association*, *280*, 1951–1958.

National Research Council (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academies Press.

Norton, I. M., & Manson, S. M. (1996). Research in American Indian and Alaska Native communities: Navigating the cultural universe of values and process. *Journal of Consulting and Clinical Psychology*, *64*, 856–860.

Office of Research Integrity (2005). http://ori.dhhs.gov/misconduct/cases/press_release_poehlman.shtml).

Oransky, M., Fisher, C. B., Mahadevan, M., & Singer, M. (in press). Barriers and opportunities for recruitment for non-intervention studies on HIV risk: Perspectives of street drug users. *Substance Use and Misuse*.

Public Health Service Act 3.01(d), 42 U.S.C. § 241(d) (1944–2006).

Rees, E., & Hardy, J. (2003). Novel consent process for research in dying patients unable to give consent. *British Medical Journal*, *327*, 198–200.

Roberts, L. W., & McAuliffe, T. L. (2006). Investigators' affirmation of ethical, safeguard, and scientific commitments in human research. *Ethics & Behavior*, *16*, 135–150.

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom (expanded ed.).*New York: Irvington.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, *5*, 127–134.

Sandler, J. C., & Russell, B. L. (2005). Faculty-student collaborations: Ethics and satisfaction in authorship credit. *Ethics & Behavior*, *15*, 65–80.

Sieber, J. E. (1991a). Openness in the social sciences: Sharing data. *Ethics & Behavior*, *1*, 69–86.

Sieber, J. E. (Ed.). (1991b). *Sharing social science data: Advantages and challenges*. Newbury Park, CA: Sage.

Sieber, J. E. (1993). Ethical considerations in planning and conducting research on human subjects. *Academic Medicine*, *9*, 59–513.

Sieber, J. E., & Levine, R. J. (2004). Informed consent and consent forms for research participants. *APS Observer*, *17*, 25–26.

Spitz, V. (2005). *Doctors from hell: The horrific account of Nazi experiments on humans*. Boulder, CO: Sentient Publications.

Swazey, J. P., Anderson, M. S., & Lewis, K. S. (1993). Ethical problems in academic research. *American Scientist*, *81*, 542–553.

Taub, H. A., Baker, M., & Sturr, J. F. (1986). Informed consent for research: Effects of readability, patient age, and education. *Law and Public Policy*, *34*, 601–606.

Varnhagen, C. K., Gushta, M., Daniels, J., Peters, T. C., Parmar, N., Law, D., Hirsch, R., Takach, B. S., & Johnson, T. (2005). How informed is online informed consent? *Ethics & Behavior*, *15*, 37–48.

Whitley, B. E., Jr. (1995). *Principles of research in behavioral science*. Mountain View, CA: Mayfield Publishing.

Whitley, B. E., Jr. & Keith-Spiegel, P. (2002). *Academic dishonesty: An educator's guide*. Mahwah, NJ: Erlbaum.

Wilson, T. D., DePaulo, B. M., Mook, D. G., & Klaaren, K. J. (1993). Scientists' evaluations of research: The biasing effects of the importance of the topic. *Psychological Science*, *4*, 322–325.

Wolf, L. E., & Lo, B. (1999). Practicing safer research: Using the law to protect sensitive research data. *IRB: A Review of Human Subjects Research*, 21, 4–7.

CHAPTER

21 Clinical Research with Culturally Diverse Populations

Frederick T. L. Leong *and* Zornitsa Kalibatseva

Abstract

The increasing population of racial and ethnic minorities calls for more attention to cultural diversity in clinical research. This chapter starts with a definition of culture and a brief discussion of the two parallel approaches to culture within psychology, cross-cultural psychology and racial and ethnic minority psychology. Subsequently, the chapter reviews cross-cultural issues in clinical research along two dimensions, namely the methodological strategies used to undertake clinical research and the methodological challenges encountered in clinical research. The reviewed methodological strategies include clinical case studies, analogue and simulation studies, randomized clinical trials, archival research and secondary data analysis, culture-specific approaches to treatment research, and meta-analysis. Lastly, the chapter discusses five methodological challenges for clinical research with culturally diverse populations, such as sample selection, measurement equivalence, race and ethnicity as demographic versus psychological variables, confounds and intersectionality, and differential research infrastructure.

Key Words: Racial and ethnic minorities, cultural diversity, methodological strategies

Cross-Cultural Issues in Clinical Research

As the number of racial and ethnic minorities and immigrants grows in the United States, so too does the need for culturally appropriate clinical research and services. From 2010 to 2050, the U.S. Census Bureau has projected that the percentage of Hispanic Americans will grow from 16 percent to 30 percent, whereas the percentage of Asian Americans will grow from 4.5 percent to 7.6 percent (U.S. Census Bureau, 2011). The percentage of African Americans will not change significantly, going from 12.2 percent to 11.8 percent, whereas the percent of White European Americans is projected to drop from 64.7 percent to 46.3 percent. Indeed, several states have become or will become "majority minority" states where traditional minority groups will outnumber the previous White European American majority group. With these important demographic changes in the U.S. population, it is not surprising that the various specialties focused on mental health (e.g., clinical psychology, psychiatry, and social work) have also increased their research and theory building to accommodate the increasing cultural diversity of the general population and, by extension, the clinical populations they serve.

Our current review of cross-cultural issues in clinical research has been preceded by other reviews (e.g., Hall, 2001; Lopez & Guarnaccia, 2000; Zane, Hall, Sue, Young, & Nunez, 2004). As such, we highlight some of the important issues raised in these previous reviews and address additional key challenges. In addition, there have been national policy reviews dating back to President Carter's Commission on Mental Health and stretching forward to President Bush's New Freedom Commission of Mental Health. Both of these commissions highlighted the problems and challenges in meeting the mental health needs of racial and ethnic minority groups in the

United States. It is perhaps the "Supplement to the Surgeon General's Report on Mental Health" (U.S. Department of Health and Human Services, 2001) that has generated the most attention in the scientific community in terms of mental health disparities for racial and ethnic minority groups. According to the Surgeon General's Report on Mental Health (U.S. Department of Health and Human Services, 1999), a range of effective, well-documented treatments exist for most mental disorders, yet nearly half of all Americans who have a severe mental illness fail to seek treatment. Whereas the Surgeon General's report did provide "hope for people with mental disorders by laying out the evidence for what can be done to prevent and treat them," its preface also noted significant gaps that "pointed out that all Americans do not share equally in the hope for recovery from mental illness" (U.S. Department of Health and Human Services, 1999). It was left to the "Supplement to the Surgeon General's Report on Culture, Race and Ethnicity in Mental Health" (U.S. Department of Health and Human Services, 2001) to document the existence of major disparities affecting mental health care of racial and ethnic minorities compared with White European Americans. The supplement highlighted that minorities have less access to, and availability of, mental health services; they are less likely to receive needed mental health services; those minority-group members who receive treatment often receive a poorer quality of mental health care; and minorities are underrepresented in mental health research (U.S. Department of Health and Human Services, 2001). The present chapter addresses some of these minority mental health issues by examining methodological strategies and challenges inherent in the conduct of clinical research with these populations.

Beginning with a definition of culture and moving on to a discussion of the two parallel approaches to culture within the psychological literature, we review cross-cultural issues in clinical research along two dimensions, namely the methodological strategies used to undertake this research and the methodological challenges encountered in such clinical research.

Two Parallel Approaches to Culture

We use Brislin's (2000) broad definition of culture to guide our review. Accordingly, "Culture refers to the shared values and concepts among people who most often speak the same language and live in proximity to each other. These values and concepts are transmitted for generations and they provide guidance for everyday behaviors" (Brislin, 2000, p. 4). As evident, culture can refer equally to racial and ethnic minority groups in the United States or to cultural groups in other countries and regions of the world. In an article recommending the integration of cross-cultural psychology research methods into ethnic minority psychology, Leong, Cheung, and Leung (2010) noted two separate and distinct disciplines that underlie the field of multicultural psychology in the United States. They cited an article by Hall and Maramba (2001), which pointed to the disconnect and lack of overlap in the literatures of these two subfields. On the one hand, the field of cross-cultural psychology has been influenced more by anthropology and cross-national studies of human behavior with a heavy emphasis on social-psychological analyses. On the other hand, racial and ethnic minority psychology has been influenced more by sociology and concerns with social stratification and social opportunities for national subgroups. Using the American Psychological Association (APA) as an illustration, Leong and colleagues (2010) noted that the latter field is represented by Division 45 (Society for the Psychological Study of Ethnic Minority Issues), whereas the former is represented by Division 52 (International Psychology). Each field has a separate history, associations, scientific journals (e.g., *Journal of Cross-Cultural Psychology* vs. *Cultural Diversity and Ethnic Minority Psychology*), and conventions and an emphasis on different philosophical orientations. Cross-cultural psychology has had a longer interest in methodological and measurement challenges, whereas racial and ethnic minority psychology has been oriented toward political advocacy, social justice, and social change. In essence, the growth in psychological research across nations parallels the development in research on cultural and ethnic groups within nations.

Although studies of the populations in their original cultures may help inform the ethnic populations in acculturated contexts in many ways (Leong et al., 2010), the focus of our current chapter is mainly within the domestic side of the multicultural psychology nexus, namely racial and ethnic minority issues in clinical research. However, as Leong and colleagues (2010) recommended, there are places where cross-pollination of ideas and methods may be useful, and that is equally true for the current chapter.

Methodological Strategies for Clinical Research

In recognizing the limitations of mono-method research, Campbell and Stanley (1966) recommended the multitrait multimethod approach.

Consistent with their concern about method bias, we have chosen to review cross-cultural clinical research by reviewing six research methods that have been used in the field. We propose that these methodological strategies have the potential to enrich clinical research with racial and ethnic minority groups. Each method will be discussed briefly and examples from the clinical literature will be provided to illustrate their potential. These methods are (1) clinical case studies, (2) analogue and simulation studies, (3) randomized clinical trials (RCTs), (4) archival research and secondary data analysis, (5) culture-specific approaches to treatment research, and (6) meta-analysis.

Clinical Case Studies

Clinical case studies provide information for examining therapy process and outcome with culturally diverse clients. These detailed accounts can illustrate important treatment issues, such as diagnosis of culture-bound syndromes, cross-cultural therapy techniques, transference and countertransference issues, therapeutic process, culturally grounded case conceptualization, and factors related to language and the use of interpreters. In addition, case studies involving therapy with culturally diverse elderly persons, couples, and families can demonstrate some of the unique challenges in providing effective treatment for these subgroups.

An underlying common theme appearing in many of these case studies is the conflict arising from differing cultural constructions of the self and interpersonal relations between the therapist and client. In particular, minority clients may have more communal and collective orientation than European American clients. At the same time, collectivist cultural values may clash with the dominant values in individual therapy that can emphasize introspection, self-awareness, and assertiveness (Constantine & Sue, 2006; Leong, Wagner, & Tata, 1995). Differences in collectivism and individualism will also depend on the cultural identity and the acculturation level of the client (Bucardo, Patterson, & Jeste, 2008). Cultural identity refers to the degree to which an individual identifies with a specific cultural group (e.g., racial or ethnic group), and acculturation refers to the process whereby members of a cultural minority group (e.g., immigrants or ethnic minorities) chose to change or not to change their behaviors and attitudes to resemble those of the host or majority cultural group. A few case studies have also addressed conflicts and issues related to ethnic identity development among racial and ethnic minority youth (e.g., Ecklund & Johnson, 2007; Henriksen & Paladino, 2009).

The evolution of the role of culture in psychopathology is partially demonstrated in the progression of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) editions. Originally, the DSM-I (APA, 1952) ignored culture and undertook an entirely universalist perspective of mental disorders. The DSM-II (APA, 1968) admitted the existence of culture but treated it as noise or a nuisance variable. In the DSM-III (APA, 1980), culture referred to remote places and exotic syndromes; the role of culture was further recognized, although it was still considered secondary. By the time the DSM-IV (APA, 1994) appeared, culture was admittedly an important dimension to assess and understand. Although some of the proposals that the National Institute of Mental Health Workgroup on Culture, Diagnosis, and Care suggested for the DSM-IV were incorporated, others were disregarded (Mezzich et al., 1999). It can be argued that future diagnostic systems need to incorporate a genuinely comprehensive framework that reflects the multifaceted nature of mental health problems. Ultimately, a last step will be to recognize culture as omnipotent and entertain the possibility that psychopathology needs to be discussed from a relativist perspective.

An important step toward the recognition of the significant role that culture plays in mental disorders was the inclusion of the Cultural Formulation approach in the DSM-IV. The Cultural Formulation model aimed at incorporating the idiographic and sociocultural factors that existing classification systems often neglect. There is increasing recognition of the value of a bidirectional approach to clinical diagnosis—top-down and bottom-up. The bottom-up approach consists of taking the perspectives of the patient and the patient's family and significant others. Therefore, focusing on the patient's personal perspective and the patient's understanding of his or her experience is fundamental. Research suggests that cultural factors are serving as moderators and mediators of clinical diagnosis, psychological assessment, and the therapeutic process (Lopez & Guarnaccia, 2000). The Cultural Formulation model in the DSM-IV (APA, 1994) includes the following sections: (1) cultural identity of the individual; (2) cultural explanation of the individual's illness; (3) cultural factors related to psychosocial environment; (4) cultural elements in the therapeutic relationship; and (5) overall cultural assessment for the diagnostic interview. Clinical case studies that

Table 21.1 Research Approaches in Clinical and Counseling Research by Degree of Control Dimensions.

		Degree of Control	
		High	**Low**
Setting	Laboratory	Experimental analogue	Correlational analogue
	Field	Experimental field	Correlational field

use the Cultural Formulation model have appeared in *Culture, Medicine, and Psychiatry*. The published case studies are useful for training purposes and encourage further research on the culture-specific elements (Bucardo et al., 2008; Cheung & Lin, 1997; Shore & Manson, 2004).

It is important to recognize the limitations of clinical case studies. While they can serve as a source of new variables for empirical operationalization and investigation, they are idiographic and limited in generalizability. Clinical case studies need to be used as only one of many methods to describe effective therapy with culturally diverse populations.

Analogue and Simulation Studies

Analogue and simulation studies in clinical research use situations that are similar to real life and provide models that are suitable for the research of psychopathology or therapy (Alloy, Abramson, Raniere, & Dyllere, 2003). These types of research strategies may be very appropriate in clinical research that seeks to understand health disparities in service utilization among minorities, therapists' biases toward minorities in therapy, and perceptions held by minority groups about mental disorders and therapy, to name a few.

Gelso (1992) presented a typology of research approaches in clinical and counseling research by combining the two levels of the Degree of Control dimension (manipulative and nonmanipulative) with the two levels of the Setting dimension (laboratory and field) to yield four types of investigative styles (Table 21.1). The first type is the experimental analogue studies, which consist of high control in a lab setting. The second type is the correlational analogue studies, which are characterized by low control in a lab setting. On the contrary, experimental field studies refer to high control in a field setting, while correlational field studies consist of low control in a field setting. Gelso (1979) proposed the "bubble hypothesis," which refers to the difficulty that one has when trying to make a bubble disappear after placing a decal on a window. In this case, the "bubble hypothesis" refers to the negotiation between internal and external validity, and it maintains that all research strategies have flaws.

Overall, field studies increase the ecological validity, whereas lab studies allow more control. In psychotherapy outcome research, the experimental field studies would be the most preferred (also known as randomized controlled trials, which will be discussed next). However, they are extremely costly and difficult to conduct, and they disrupt normal clinical services in treatment agencies; hence, there is great resistance among these agencies to permit such studies. By contrast, the correlation field studies tend to be a weaker design, since the lack of control and manipulation of variables in the field tends to yield findings with limited causal inferences.

According to Gelso (1992), analogue research in therapy most often involves studies of intervention components. The intervention is not studied as it actually is or naturally occurs, but instead as it is approximated by the researcher. Also, the typical analogue examines some aspect of the intervention rather than the entire intervention.

Among the many aspects of therapy, one can study the effects of (1) therapist techniques, behavior, personality, appearance, intentions, or style; (2) numerous client characteristics; and (3) different kinds or characteristics of treatments. These studies enable us to examine how specific independent variables affect some aspect of the client's behavior, some aspect of the therapist's behavior, some aspect of their interaction, or all of these.

The major advantage of the experimental analogue is that it has high internal validity. By tightly controlling and manipulating the independent variable, it allows for strong causal inferences. On the other hand, its disadvantage is lower external validity (ecological validity). By controlling and isolating the effects of a single independent variable, that variable must be pulled out of its natural context and less represents the complex nature of therapy as it occurs.

One can describe two basic kinds of analogues: the audiovisual analogue and the quasi-intervention

analogue. In the audiovisual analogue, the subject is presented with a stimulus tape or film (although at times only written stimuli are used), asked to assume the role of the therapist or the client, and asked for different responses at various points in time. The tape or film usually displays a client, a therapist, or both. The researcher has control over what transpires on the tape or film.

In the quasi-intervention analogue, one or more interviews are held, and the activities that occur may vary greatly in how closely they approximate the actual intervention being studied (Gelso, 1992). Typically, the following characteristics are displayed: (a) one or more interviews are held in which (b) therapy is approximated to one degree or another, and (c) a confederate client or therapist exhibits behaviors or characteristics that are prearranged or predetermined by the experimenter and are (d) systematically varied in order to (e) assess their effects on certain behaviors of the client (participant), the therapist (participant), or their interaction.

To illustrate how analogue and simulation studies may be useful for the clinical research with racial and ethnic minorities, we provide two examples. In an analogue study, Li, Kim, and O'Brian (2007) examined the effects of Asian cultural values and clinician multicultural competence on the therapy process. The participants were 116 Asian American college students who watched analogue videotapes of a European American female "therapist" seeing an Asian American female "client." If the "therapist" expressed cultural values inconsistent with Asian culture, she was considered more culturally competent by participants when she acknowledged racial differences between herself and the client. In addition, when the Asian American observer-participants adhered strongly to the Asian value dimension of conforming to norms, they also tended to perceive the "therapist" as more credible and culturally competent.

In a simulation study, Fernandez, Boccaccini, and Noland (2008) assessed the validity of the English and Spanish versions of the Personality Assessment Inventory (PAI) and their ability to detect simulated responding. The study included 72 bilingual students and community members who filled out both the English- and Spanish-language versions of the PAI. The participants were instructed to respond honestly, to overreport psychopathology for an insanity case, or to underreport psychopathology for an employment evaluation. The authors found that the validity scales in both English and Spanish performed similarly.

Randomized Control Trials

Randomized controlled trials (RCTs; see Chapter 4 in this volume) serve as the gold standard for establishing the effectiveness of treatment. RCTs represent the most rigorous method for evaluating the effects of various treatments, since the treatments are randomly administered to patients to reduce confounds, are controlled by using single-, double-, or triple-blind procedures, and are applied to various control groups.

Whereas RCTs are methodologically sound, their usefulness can be limited by ethical and practical concerns. For true randomization, researchers and clinicians assign their participants to intervention without consideration or speculation about which is better (see Chapter 4 in this volume for further consideration of ethical issues related to RCTs). The idea is that the evaluation via RCT will inform the field about the relative effectiveness of different therapies. For example, Rossello, Bernal, and Rivera-Medina (2008) conducted a study that compared individual and group formats of cognitive-behavioral therapy (CBT) and interpersonal therapy for depressed Puerto Rican adolescents. The participants (n = 112) were randomized to four conditions (individual or group CBT; individual or group interpersonal therapy). The results of this study suggested that both CBT and interpersonal therapy are effective treatments for depression among Puerto Rican adolescents, although participants who received CBT reported greater decreases in depressive symptoms. This RCT is one of very few that have been conducted with culturally diverse populations (e.g., Huey & Polo, 2008; Miranda et al., 2005).

Although RCT is a sound methodology, it may be too stringent a criterion for cross-cultural interventions in practice because of the differential infrastructure problem (see below; Leong & Kalibatseva, 2010). Despite the advance of cross-cultural psychotherapy research in the past decade, many of the currently used culturally sensitive treatments have not been subjected to RCTs. However, it is not practical to ask mental health professionals to stop conducting therapy with culturally diverse populations and to wait for RCTs to be conducted.

Various limitations and advantages are associated with RCTs. One advantage is that RCTs can be used to compare multiple active treatments—using a rigorous experimental design without withholding treatment deemed necessary for a clinical population (see Chapter 4 in this volume). RCTs are usually more expensive and time-consuming to

conduct than other studies. However, as noted earlier, RCTs adhere to the research strategies, which Gelso (1992) framed as experimental field studies, or actual clinical cases in treatment agencies, which are characterized by high external and internal validity.

RCTs have played an important role in clinical psychology research. In 1993, a Division 12 (Society of Clinical Psychology) Task Force established criteria for empirically validated treatments. The task force members recommended that the field systematically review and classify therapeutic interventions in two categories: well-established treatments or probably efficacious treatments. Following discussions, the field moved to the concept of empirically supported therapies, which suggests that different therapies may have varying degrees of support. Chambless and Hollon (1998) defined empirically supported therapies as treatments that have been demonstrated to be superior in efficacy to a placebo or another treatment. In anticipation of more criticism, Chambless and Hollon acknowledged that some clinicians and researchers would disagree that RCTs and single case experiments were the best means to detect causality and that evidence of efficacy precedes effectiveness.

This notion of empirically supported therapies has also evolved, with the field referring to the broader concept of evidence-based practice, as used in medicine. The foundations of evidence-based practice were laid with Cochrane's (1979) report, which argued for assembling and renewing periodically all scientific evidence related to treatment approaches that have proven to be effective using RCTs. The Cochrane Collaboration (http://www.cochrane.org/reviews/clibintro.htm), established in 1993, has served as the exemplar of evidence-based practice.

CROSS-CULTURAL COMPETENCIES IN THERAPY

At APA, Division 12, Section VI (Clinical Psychology of Ethnic Minorities), was established to promote research on clinical interventions with racial and ethnic minorities; to encourage sensitivity to cultural, ethnic, and racial issues in training; to foster training opportunities for racial and ethnic minority clinical psychologists; and to promote communication about sociocultural issues within clinical psychology (http://www.apa.org/divisions/div12/sections/section6/). A task force of Counseling Psychology (17) and the Society for the Psychological Study of Ethnic Minority Issues (45) developed guidelines on multicultural competencies (APA, 2003). Despite their impact, a lingering issue has remained: the "criterion problem." This issue concerns the lack of research evidence that a culturally competent therapist produces better client outcomes than a therapist who is not deemed culturally competent. It seems reasonable that the evidence-based practice approach could be adopted. As suggested by Cochrane (1979), we need to be guided by a critical summary of the best available scientific evidence for how we approach our practice. But the question is, how do we reconcile these two movements? What do you do when there are no studies to provide the evidence for psychotherapy with racial and ethnic minority patients? Undoubtedly, withholding treatments for these patients until evidence can be accumulated cannot be the solution.

Efficacy involves clinical research in controlled laboratory settings, whereas effectiveness involves the applications of efficacious treatments in actual clinical settings in which there is much less experimental control (Hall & Eap, 2007). In the field of therapy, there is less research available regarding treatment effectiveness than treatment efficacy, a situation that can be seen as problematic for cross-cultural therapy because both effectiveness and efficacy studies are extremely scarce for racial and ethnic minority groups. Nevertheless, the concept of evidence-based practice in clinical psychology has become the standard as both federal agencies and training programs follow this model.

In the APA 2005 Presidential Task Force on Evidence-Based Practice, the issue of treatment for racial and ethnic minority groups was addressed. The report noted that client characteristics such as age, culture, race, ethnicity, gender, gender identity, religious beliefs, family context, and sexual orientation need special attention (APA, 2006). All of these attributes influence the client's "personality, values, worldviews, relationships, psychopathology, and attitudes toward treatment" (p. 279). Culture influences the nature and expression of psychopathology as well as the explanatory models of health and illness, help-seeking behaviors, and expectations about treatment and outcomes (APA, 2006; Lopez & Guarnaccia, 2000). The report noted that future research needs to address a myriad of issues, including the effect of patient characteristics on seeking treatment, treatment process, and outcomes, as well as empirical evidence for the effectiveness of psychological interventions with racial and ethnic minorities. In addition, the report made an important observation: "Evidence-based practice in psychology (EBPP) is the integration of the best available

research with clinical expertise in the context of patient characteristics, culture, and preferences" (p. 273). Since our knowledge of the influence of patient characteristics, culture, and preferences among minorities in psychological assessment, case formulation, therapeutic relationship, and interventions is limited, future clinical research needs to explore them within the framework of evidence-based practice in psychology. Ultimately, the goal of this research should be to answer the question, "What works for whom, when, and under what conditions?"

Archival and Secondary Data Research

Given that RCTs are very expensive and difficult to carry out, another method for conducting clinical research with minority populations is to use archival datasets or secondary data analysis. Major advantages of archival data include the time and cost savings involved in direct data collection; it is also an unobtrusive method that does not interfere with the natural procedures within clinical centers and hospitals (sometimes referred to as naturalistic research). Archival data analysis is also particularly helpful when the researchers do not have direct access to a large number of minority-group members (e.g., studying small ethnic groups in particular geographical areas, such as Asian Americans in the Midwest). Another major advantage of using archival data is that it is high on external or ecological validity because researchers can investigate naturally occurring and ongoing clinical phenomena (as opposed to phenomena simulated in a lab setting).

A major disadvantage, however, is the fact that research questions are limited to the variables and samples that have already been collected. Using Gelso's "bubble hypothesis," this method is high on external validity but sacrifices internal validity and the ability to make causal inferences. Studies based on archival datasets tend to be correlational field studies since there is no control or manipulation of independent variables or random assignment of participants or patients. Next, we provide an example of a study that used archival data.

Sue and colleagues used archival data from the automated information system maintained by the Los Angeles County Department of Mental Health to examine the effects of ethnic matching between therapist and client on the type, length, and outcome of outpatient services (Sue, Fujino, Hu, Takeuchi, & Zane, 1991). The archival data were collected for the purpose of system management, revenue collection, clinical management, and research. The data consisted of client, therapist, and treatment variables and were collected on 600,000 different clients who received outpatient mental health services between 1973 and 1988.

The study tested the "cultural responsiveness hypothesis," which assumed that therapist–client matching in terms of ethnicity and language would result in more advantageous outcomes for clients. The participants in this study were African Americans, Asian Americans, Mexican Americans, and Whites who received services in the past 5 years. Sue and colleagues (1991) found partial support for the cultural responsiveness hypothesis, as clients in ethnically matched therapeutic dyads had longer treatments; only Mexican Americans reported better outcomes when they were ethnically matched. At the same time, clients whose primary language was not English and who were matched by language and ethnicity stayed in treatment and reported better outcomes. This study is one of the first archival data studies that examined mental health service use among large samples of ethnic and racial minority clients. While using archival data may not allow researchers to manipulate any variables, it provides numerous advantages, such as large samples, already collected data, and generalizability of findings.

Secondary data analysis refers to using data that have already been collected by someone else, and it has several advantages (Hofferth, 2005). First, secondary data analysis is both economical and time-saving because the researcher does not have to expend the resources or the time to collect the data. Second, representative sampling techniques are often used for data collection. Third, the datasets often have large sample sizes and high response rates of groups that may be difficult to reach and sample. However, Hofferth also warned researchers of some of the disadvantages of secondary data analysis, such as measurement issues, concerns with "fishing" the data, considerable investment of time to learn how to analyze the data, and restriction to a specific set of variables that may not answer the particular research questions.

Culture-Specific Approaches to Treatment Research

With increasing diversity in society, researchers and clinicians recognize the need to consider the role of culture in therapy. We examine two different culture-specific approaches to treatment research that have been developed and researched: (1) the Cultural Accommodation model and (2) the Cultural Adaptations of Therapies approach. The

major difference between them is that the Cultural Accommodation model assumes that the therapist will accommodate to the cultural background of the client, whereas the culturally adapted therapies try to incorporate cultural modifications into treatment itself.

The development and refinement of the Cultural Accommodation model transpired over a long period of time, and the final goal of this process was the formulation of an integrative and multidimensional model of cross-cultural psychotherapy (Leong, 2007). Drawing on Kluckhohn and Murray (1950), and the oft-cited quote "Every man is in certain respects: a) like all other men, b) like some other men, and c) like no other man" (p. 35), Leong (1996) developed a multifaceted cross-cultural model of psychotherapy. Accordingly, there are personality determinants in people's biological and genetic makeup, which represent the Universal dimension of human identity (aka biological aspect of the biopsychosocial model). Other personality characteristics are observed at the Group dimension, as most men are like some others, with social groupings based on ethnicity, race, culture, or social class (Leong, 2007). Kluckhohn and Murray also noted that each person is individual and unique in his or her perceptions, feelings, needs, and behaviors. This idea captures the Individual dimension of human identity and emphasizes the distinct social learning experiences, values, beliefs, and cognitive schemas of each person.

In the pursuit of universal laws of behavior, important factors such as ethnicity and gender may be downplayed. The fields of gender psychology and ethnic minority psychology deal directly with cross-gender and cross-cultural issues. At the same time, referring to people at the group level can unwittingly encourage stereotyping. Clinical psychologists are reminded that every person has unique individual experiences that cannot be explained by the group or universal levels (Leong, 2007).

Leong (1996) stresses that mental health professionals need to address the Individual, the Group, and the Universal. The Individual dimension concerns unique characteristics and differences that have been studied by behavioral and existential theories—in particular, individual learning histories and personal phenomenology centered on Individual variations in order to understand human behavior and psychopathology. The Group dimension has been the center of attention for cross-cultural psychology and psychopathology, ethnic minority psychology, and gender psychology. Finally, the Universal dimension covers mainstream psychology and the "universal laws" about human behavior and psychopathology that have been established (e.g., the "flight-and-flight" response in cases of threat has been considered universal for humans).

Leong and Tang (2002) insisted that concentrating only on the Universal dimension disregards the Individual and Group dimensions that are essential parts of human behavior. In an integrative model, the Universal dimension has a vital role and can explain some behaviors, but it is not comprehensive enough (Leong, 1996). According to Leong (1996), the Group dimension is equally central in explaining human personality and the groupings may vary from culture, race, and ethnicity to social class, occupation, and gender. Members who belong to the same group share a bond and are different from members from other groups. Lastly, the Individual dimension of human personality stands for the uniqueness of each person, because no two individuals are identical in everything. If we ignore the Individual dimension, we risk stereotyping persons from different cultural groups and overgeneralizing the Group dimension.

In psychotherapy, there may be more emphasis on the Universal dimension and the client may be more closely associated with the Group dimension. To form a therapeutic alliance with the client, the therapist may need to move toward the Group level while trying to steer away from stereotyping the person and attending to the Individual dimension. Careful assessment of all three levels is essential to understanding the client and coming up with an accurate formulation of the case (Fig. 21.1).

Leong and Serafica (2001) argued that there are three major approaches to cultural differences within psychology. First, the authors described the *universalist* approach, which considers culture as noise or a trivial variable that requires little or no attention in clinical research and practice. Mainstream Western psychology has historically favored this approach, and it is represented in the Universal dimension of Leong's (1996) integrative psychotherapy model. Second, Leong and Serafica described the *culture assimilation* approach, which recognizes existing cultural differences but assumes that ethnic and racial minorities should assimilate to the mainstream U.S. culture (i.e., Western European culture). The main idea behind the assimilation is that the existing psychological theories and models will work for the culturally diverse groups once they become part of the "melting pot." Third, Leong and Serafica proposed the *culture accommodation* approach, which

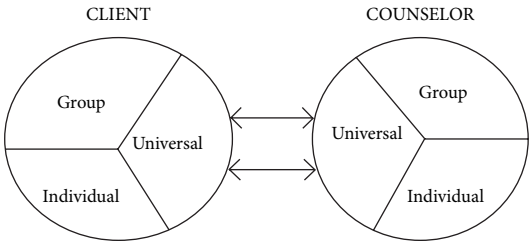


Figure 21.1 Multidimensional model: An integrative tripartite model of cross-cultural counseling.

considers culturally unique experiences of minority groups as significant factors in understanding their behavior. Once these culture-specific factors are identified, they are included in current theories and models to enhance their relevance and utility to minority groups.

According to Leong (2007), the Cultural Accommodation model recognizes the importance of using in conjunction the Universal, Group, and Individual dimensions when working with culturally diverse populations. The cultural gaps and continuing cross-cultural issues within our Western models of psychopathology and psychotherapy are evidenced in the twin enduring problems of underutilization of mental health services and premature termination from therapy among minority groups. The goal of the Cultural Accommodation model is to utilize existing theories and models and to incorporate culture-specific variables into the assessment process and theory formulations in order to make our psychotherapeutic interventions more effective and culturally relevant (Leong & Tang, 2002).

Leong's (1996) integrative Cultural Accommodation model incorporates the Universal, Group, and Individual dimensions of human personality and may improve the cultural validity of existing models of psychotherapy. Moreover, the Cultural Accommodation model uses a person–culture interaction model, which allows researchers and clinicians to focus on the cultural context variables. A few of the variables that may be important to consider and measure in clinical research and practice are cultural identity, acculturation level, loss of face, interpersonal relatedness, and collectivism.

The Cultural Accommodation model proposes to examine the cultural validity of the current models of psychotherapy and to pinpoint culturally relevant variables that will address the deficiencies of existing models and increase their effectiveness with minority clients. To evaluate the cultural validity of our models and justify the choice of culture-specific variables, Leong and Lee (2006) proposed applying the evidence-based practice approach. The authors suggest that the psychopathology or psychotherapy model or theory should be examined in detail and the culture-general aspects should be narrowed down. Elements specific to Western European culture should not be generalized to other cultures and forced upon minority groups. Leong and Lee (2006) demonstrated the use of the Cultural Accommodation model with Asian Americans. Some of the relevant culture-specific constructs that they found for this group were cultural identity, acculturation, self-construal, individualism–collectivism, and high–low context communication styles. Each one of these variables may moderate or mediate the therapeutic process and, therefore, needs to be measured and accounted for in providing effective psychotherapy.

Hall and colleagues (Hall, Hong, Zane, & Meyer, 2011) presented mindfulness and acceptance-based psychotherapies as promising treatment modalities for Asian Americans. For instance, they identified two aspects of contextual therapies that emphasize primarily a Western worldview: "1) the marked distinction between the self and others ('I' vs. 'You'); and 2) the importance of active coping by the self" (p. 219). Specifically, these features of mindfulness and Acceptance and Commitment Therapy pertain to differences in self-construal (interdependent vs. independent) and styles of coping (implicit vs. active). Hall and colleagues noted that the notion of the transcendent self in Acceptance and Commitment Therapy concentrates entirely on the importance of the self as opposed to others, which may run counter to the importance of interdependence among Asian Americans. Similarly, the Western idea that one needs to actively cope with feelings, wants, and values as opposed to suppress them is also contradictory to Asian Americans' notion of preserving group harmony (Leong & Kalibatseva, 2011). Therefore, Hall and colleagues propose cultural adaptations for mindfulness and Acceptance and Commitment Therapy when used with Asian Americans.

In summary, the Cultural Accommodation model of psychotherapy (Leong & Lee, 2006) entails three stages: (1) identifying cultural blind spots in current theories and models that impede cultural validity; (2) choosing culture-specific constructs that are supported by existing research in the cross-cultural and racial and ethnic minority literature as important; and (3) testing the incremental validity of the culturally accommodated theory.

One parallel development to the Cultural Accommodation model is the Cultural Adaptations of Therapies approach, which aims at integrating the cultural competence literature with that of evidence-based practice in psychology (e.g., Hwang, 2006; Lau, 2006). The cultural adaptation models resemble the Cultural Accommodation model in that they detect important cultural variables to integrate into the evidence-based practice of psychotherapy and provide further convergent evidence. While the Cultural Accommodation model concentrates mainly on the therapist accommodating to the cultural background of his or her clientele, the culturally adapted therapies modify the treatment approaches. Nonetheless, both approaches promote the goal of providing culturally relevant and effective mental health services (Leong & Kalibatseva, 2010).

Culturally adapted or culturally sensitive treatments entail "the tailoring of psychotherapy to specific cultural contexts" (Hall, 2001). Whaley and Davis (2007) define cultural adaptation as "any modification to an evidence-based treatment that involves changes in the approach to service delivery, in the nature of the therapeutic relationship, or in components of the treatment itself to accommodate the cultural beliefs, attitudes, and behaviors of the target population" (pp. 570–571). Evidence-based treatments or empirically supported treatments refer to "the interventions or techniques (e.g., cognitive-behavioral therapy for depression, exposure therapy for anxiety) that have produced therapeutic change in controlled trials" (Kazdin, 2008). Since the majority of evidence-based treatments have historically been developed and tested primarily with Caucasian Americans, mental health professionals have questioned their efficacy with minority populations. According to Whaley and Davis, while some researchers believed that including minority patients in efficacy trials is a sufficient form of cultural adaptation (e.g., Chambless et al., 1996), others argued that substantial cultural adaptations are necessary in the approach, delivery, therapeutic process, and inclusion of cultural knowledge (regarding attitudes, values, beliefs, and behaviors) to have more culturally appropriate empirically supported treatments (Atkinson, Bui, & Mori, 2001; Miranda, Nakamura, & Bernal, 2003; Muñoz & Mendelson, 2005). In the past decade, there has been a growing interest in adapting existing empirically supported treatments for various minority groups, such as behavioral activation for depressed Latinos (Kanter, Santiago-Rivera, Rusch, Busch, & West, 2010), group CBT for depressed African Americans (Kohn, Oden, Muñoz, Robinson, & Leavitt, 2002), group and individual CBT and interpersonal therapy for Puerto Rican adolescents (Rossello, Bernal, & Rivera, 2008), trauma-focused CBT for American Indian and Alaskan Native children (BigFoot & Schmidt, 2010), exposure-based CBT of anxiety for Latino youth (Pina, Villalta, & Zerr, 2009), and exposure treatment for phobic Asian Americans (Pan, Huey, & Hernandez, 2011).

Kohn and colleagues (2002) described a culturally adapted group CBT for depressed African American low-income women. Some of the structural changes they made included participation of only African American women, closing the group to new participants to facilitate cohesion, adding experiential meditative exercises (i.e., relaxation) and a termination ritual, and changing the language (e.g., therapeutic exercise instead of homework) to make it easier for the women to identify with it. The process changes included four modules: deconstructing the "Black superwoman" myth; exploring spirituality and religiosity; reinforcing the importance of family; and discussing African American female identity. Participants were given the opportunity to choose between the regular CBT and the adapted CBT for African American women, and 83 percent (10/12) chose the adapted one. Of those 10 women, 8 completed therapy and were compared to 10 demographically matched women from previous regular CBT groups. The mean age of the participants in the sample was 47 years. The intervention consisted of 16 weekly 90-minute group therapy sessions. Women in the culturally adapted CBT group exhibited a decrease in depressive symptoms from the severe range to the moderate range. When compared with demographically matched women, the decrease of depressive symptoms in the culturally adapted group was twice that in the regular CBT group (–12.6 vs. –5.9 points on the Beck Depression Inventory).

Videotapes of each session were coded and the predominant affective tone of the sessions with African American women was identified as intense irritability rather than sadness or anhedonia. Kohn and colleagues found this was consistent with previous research and clinical observations. Overall, the authors concluded that this pilot study had positive results, but women still needed further treatment to reduce their depressive symptoms and distress. Yet, based on the coded videotapes, the manifestation of depression among African American women may differ from what has been established to be depression among Caucasian Americans.

Evidence that Cultural Adaptation of Therapies is becoming an important approach to cross-cultural psychotherapy research will be provided later when we review meta-analysis. Griner and Smith (2006) and, more recently, Smith, Domenech Rodrigues, and Bernal (2011) provided a meta-analysis of culturally adapted mental health interventions with racial and ethnic groups. More and more psychotherapy research will be conducted using both of these approaches of cultural accommodation and cultural adaptation of therapies.

Meta-Analysis

Meta-analysis represents the last of the methods we will discuss for clinical research. As a method for accumulating the results of empirical studies and providing a quantitative index (effect size) of the evidence for or against a particular intervention or treatment, it is a valuable tool and fits well with the current movement toward evidence-based practice in psychology (see Chapter 17 in this volume for a complete discussion of meta-analysis). However, the results of a meta-analysis are only as good as the studies on which it is based, and it is of limited value in understanding the outcomes of psychotherapy for culturally diverse groups when there is a lack of representative studies, an insufficient number of representative studies, or poorly conducted studies of relevance, as delineated by Zane, Hall, Sue, Young, and Nunez (2004). In 2006, *Psychotherapy: Theory, Research, Practice, Training* published a special issue on "Culture, Race and Ethnicity in Psychotherapy." As the guest editors of this special issue, Leong and Lopez (2006) pointed out that the articles took significant steps toward bridging the fields of cultural competence and empirically supported treatments. Furthermore, they emphasized the need for "empirical tests of various issues raised, conceptual models, cultural adaptations, and correlates of therapists' multicultural competence" (p. 379).

The special issue included Griner and Smith's (2006) meta-analysis of evidence-based culturally adapted mental health interventions. The meta-analysis of 76 studies found an average treatment effect size (d = .45) from before to after the intervention, which suggested a moderately strong benefit of culturally adapted interventions. In addition, they found that treatments for groups of same-race participants were four times more effective (d = .49) than treatments for groups of mixed-race participants (d = .12). This finding suggests that cultural adaptations for specific groups may be more beneficial than general multicultural adaptations. Another important finding was that effect sizes of culturally adapted treatments increased when participants were older and when there was a higher percentage of Hispanic participants. The authors attributed the greater benefits of cultural adaptations for these populations to the impact of acculturation, suggesting that older populations may be less acculturated than younger populations, and some Hispanic populations who do not speak English may be less acculturated. In addition, when the therapist spoke the participant's native language (if not English), the treatment effect was better (d = .49) than when the therapist did not speak the participant's native language (d = .12). This meta-analysis was a great advance in the exploration of culturally adapted interventions. A more recent meta-analysis (Smith et al., 2011) reported a similar moderate effect size (d = .46) and pointed out that the most effective treatments tended to be those with a greater number of cultural adaptations. A logical next step is to review the nature of the cultural adaptations and to test whether they contribute to the already existing treatments.

Methodological Challenges for Clinical Research

As mentioned above, in using the various methodological strategies for clinical research with racial and ethnic minority groups, there is an accompanying set of methodological challenges that require attention. We will discuss five of these major methodological challenges for clinical research: (a) sample selection, (b) measurement equivalence, (c) race and ethnicity as a demographic versus psychological variable, (d) from confounds to intersectionality, and (e) differential research infrastructure. This is by no means an exhaustive list of methodological challenges; instead, it presents a set of issues that we believe are significant at present.

Sample Selection

A major methodological challenge in cross-cultural clinical research is sample selection. Sue, Kurasaki, and Srinivasan (1999) noted that the research principles of selection and sampling of the population are no different for cross-cultural or ethnic population research than for research in the general population. However, they pointed out that selection and sampling a population are more complicated when conducting research on ethnically diverse populations. They highlighted the challenges of finding representative and adequate sample sizes when conducting research with racial and ethnic

minority groups, which they attributed to the sensitive nature of the topic under study, the unfamiliarity of the respondents with the research process, the relatively small size of some ethnic populations, and cross-cultural communication issues. This situation may make it difficult to satisfy traditional research criteria, which call for random sampling to achieve a representative sample that allows greatest generalizability of the findings.

Although the situation is changing, with some states becoming "minority majority states" as noted above, finding adequate samples for cross-cultural clinical research remains a challenge. Although obtaining adequate samples of African Americans may be difficult in some states, it is almost impossible to get sufficient numbers of Chinese Americans in any state due to their extremely low base-rates. In those cases, researchers are basically attempting to sample rare events. Sue and colleagues (1999) noted that the problems of sampling rare events and finding adequate sample sizes have caused some researchers to collapse ethnic categories among the races, which in turn limits the interpretation and generalizability of the findings. They further noted that collapsing across groups is common in studies of Asian and Pacific Islander Americans, which are highly heterogeneous and diverse groups of people from Asia and the Pacific. The authors warned: "By considering the Asian and Pacific Islander Americans as a homogeneous group, we ignore sociohistorical, cultural, economic, and political diversity." (Sue et al., 1999, p. 62). Trimble (2005) proposed the term "ethnic gloss" to represent this tendency to lump very heterogeneous ethnic groups together for the sake of expediency and simplicity. This is problematic for a number of groups, including different American Indian tribes, various groups of Latinos, Asian-Pacific Islander Americans, and even White European Americans.

Small subpopulations also encourage researchers to resort to using convenience samples when conducting research with racial and ethnic minority groups (Sue et al., 1999). These samples suffer from lack of representativeness, which in turn restricts generalizability. Such overdependence on convenience samples can have larger implications. For example, they can significantly skew the results of meta-analyses that examine the cumulative effects of research in a particular domain. It may be useful for such meta-analytic studies to monitor, compute, and address the issue of convenience. In addition, existing studies with small samples (associated with wide standard deviations) may provide unstable findings.

Owing to the difficulties of obtaining adequate and representative samples for cross-cultural clinical research, the use of secondary analysis of archival data may be increasing in frequency. Secondary analysis has the advantage of using large samples and provides more stable estimates. At the same time, the use of complex sampling and weighting procedures allows researchers to estimate population estimates with these datasets. The American Psychological Association has recently published a book on the method of secondary data analysis (Trzesniewski, Donnellan, & Lucas, 2010), which indicates the heightened interest of researchers in this methodological strategy.

Measurement Equivalence

Most approaches to therapy consider diagnosis and assessment an important first step in the therapy process. Indeed, it has been argued that clinical diagnosis is of great importance because appropriate treatment depends on the correct diagnosis (Garfield, 1984). Yet, because of a significant number of problems associated with the clinical diagnosis of psychopathology, the value of the diagnostic process has been questioned (Garfield, 1984; Matarazzo, 1978). Any lack of validity in clinical diagnosis has important consequences. Errors in diagnosis have costs associated with them, in the form of either not receiving available treatment or receiving inappropriate treatment. Rosenhan's (1973) famous study with pseudopatients illustrating the lack of validity in clinical diagnosis also demonstrates the costs of diagnostic errors. These pseudopatients were kept in psychiatric hospitals for 7 to 52 days after totally relinquishing their "symptoms." The problems with diagnosis are further complicated by the cultural background of patients and therapists. Despite these problems, the clinical diagnosis of psychological problems remains an important part of psychotherapy since, when performed reliably and accurately, it can serve as a valuable guide to treatment.

In addition to diagnosis, clinical psychology relies a great deal on psychological tests to evaluate patients and treatment outcomes. The reliability and validity of our clinical measures are, therefore, fundamental to good clinical science and practice. Among the methodological problems in cross-cultural research, a fundamental issue concerns the measurement equivalence of tests and measures when used with racial and ethnic minority patients. Without evidence of measurement equivalence, the findings of cross-cultural and ethnic minority research may remain suspect. The establishment of

measurement equivalence is a fundamental requirement in cross-cultural research and should be so for racial and ethnic minority psychology.

Based on Berry's (1980) critical review of measurement equivalence issues in cross-cultural research, Leong and colleagues (2010) recommended that cross-cultural researchers attend to and evaluate the measurement equivalence of their tests and measures along the four major types: linguistic, functional, conceptual, and metric. *Linguistic equivalence* (or translation equivalence) is concerned with whether the words carry the same meaning and are associated with the same referent across cultures (e.g., ishin-denshin, which refers to tacit understanding). *Functional equivalence* may be established by showing that two or more behaviors in different cultures are functionally related to similar problems (e.g., marriage). *Conceptual equivalence* may be established by a common set of behaviors that define a construct (e.g., psychotherapy or posse). *Metric equivalence* may be shown if psychometric properties of two sets of data are the same for different cultural groups (e.g., response bias, factorial invariance).

As Leong and colleagues (2010) pointed out, simple translation is not an adequate procedure when transporting the use of clinical measures from one linguistic group to another. Since linguistic equivalence is mainly concerned with translating psychological measures from one language into another for use in another culture, what is the procedure of choice? Over the years, the back-translation method has become the standard procedure and technique of choice in cross-cultural research (Brislin, 1970, 1980). The method involves translating a scale into a different language and then translating it back into the original language. The back-translated version is checked against the original version, and problems of inaccuracy and distorted meaning are then resolved. Whereas this has been an issue in international and cross-cultural research, linguistic equivalence and translational problems have not received much attention in racial and ethnic minority research. Nevertheless, the lack of linguistic equivalence creates the same set of problems for the issue of psychological tests and measures for racial and ethnic minority groups as it does for individuals in other countries and cultures.

Classical test theory has been the main analysis framework used in measurement research for decades. However, this approach generally fails to account for such things as differential interpretation of test items across cultures and ethnic or gender differences in how individuals respond to certain instruments. Researchers have often ignored these issues in their work, but there are now various methods for addressing them. Hence, metric equivalence has become a major focus in cross-cultural measurement research in recent years. Classical test theory divides an observed score into a true score component and the error inherent in measurement. Differences between scores represent both actual differences on a certain construct and nonsystematic measurement error. The construct of interest is expected to cause variations in observed scores. Researchers often compare means of composite scores and rely on evidence of validity and reliability to justify its use across groups, but they are still assuming that the construct has conceptual equivalence across groups, that the measurement error factors are equivalent, and that observed scores relate to the intended construct in the same way. These assumptions are rarely tested, but if they do not hold, the lack of equivalence will jeopardize the validity of the conclusions. Lately, item-response theory (see Chapter 18 in this volume) has been used to examine differential item functioning. Differential item functioning refers to the case in which persons of equal ability level (e.g., equal level of distress) have a different probability of answering an item "correctly" (e.g., endorsing a symptom). When differential item functioning is present, members of different demographic groups may be more (or less) likely to endorse a test item, which suggests that the test or measure does not have the same implications for both groups.

A useful resource is Vandenberg and Lance's (2000) systematic review of developments in this area. To address the challenges of evaluating and demonstrating metric equivalence, the recommended tests generally fall into two categories. Tests for invariant covariance matrices, configural invariance, metric invariance, scalar equivalence, and invariance of unique variances of items can all be classified as tests of *measurement invariance*. Tests of invariant factor variances and covariances, and factor means can be classified as tests of *structural invariance*. Tests in the former category deal with variables' relationships to latent constructs, while those in the latter category deal with the actual latent variables of interest. These tests have been used in a number of areas, including examination of test administration methods, cross-cultural generalizability, and longitudinal study of human development (Vandenberg & Lance, 2000). Tests vary in the frequency and order with which they are used across studies. It is possible that a lack of knowledge

of the procedures or a lack of appropriate software contributed to the absence of the application of some of these tests.

Vandenberg and Lance's (2000) article revealed that there is some disagreement in the literature about the terminology, sequencing, and appropriateness of many of these tests. Most authors agree that the equality of the covariance matrices should be tested first; if these matrices are invariant, equivalence is established and further testing is unnecessary, but if they are not, further tests need to be conducted to identify the source of variance. Configural invariance should be tested second, both because it serves as a baseline for further tests and because this variance indicates the measurement of different constructs across groups. This type of testing examines factor loadings on the items across groups. We note that there is no clear agreement on the proper order of testing after these two steps. We also note that although there is general agreement on covariant matrix analysis as a first step, very few authors actually do this step, choosing instead to focus on more specific tests first.

Based on Vandenberg and Lance's (2000) review, Byrne and colleagues (2009) have outlined the steps to be taken in evaluating invariance. First, the covariance matrices are examined. Second, the patterns of factor loadings are examined to evaluate configural equivalence. In the third step, metric equivalence is investigated by testing the values of factor loadings across groups. Fourth, scalar invariance is evaluated by examining differences in the intercepts of items. Fifth, the uniqueness of each variable is investigated, although this equivalence is often not achieved and often not of interest to the researcher. Sixth, factor variances across groups are compared to see if they are equal. Seventh, factor covariances are compared to determine equality across group. Lastly, factor means are tested for equivalence across groups. Not all of these tests are needed for every study, and tests of partial invariance may be used on a subset of subgroup parameters when variance is found in one of the eight steps. More recently, Schmitt and Kuljanin (2008) provided an update to Vandenberg and Lance (2000) and identified five levels in which measurement invariance could be assessed: item-level, scalar, factorial, partial, and unique variance. While all five levels are important, the discovery of partial invariance and unique variance appears to be of greatest relevance to cross-cultural research for identifying important and systematic cultural differences.

Within Berry's (1980) four-dimensional model of measurement equivalence, conceptual equivalence is perhaps the most complex to evaluate and establish. This may account for the lack of a standard procedure or consensus regarding the best approach to conceptual equivalence in contrast to the back-translation method for linguistic equivalence. A crude but common approach to conceptual equivalence has been for researchers to use regression methods to evaluate whether regression parameters of the criteria are similar across cultural groups. Although this is economical and simple, the differences in response variability and measure reliability across cultures may lead to fluctuations in parameters that are difficult to disentangle.

Usunier (1998) has recommended one particular approach that uses decentering in the process by relying on multiple source and target languages. Researchers start with a broad, nearly etic conceptual area that is believed to be applicable to nearly all cultures. Next, native speakers of the cultures of interest are invited to generate a list of words that relate to this conceptual area. Subsequently, a cross-cultural research team works together to identify the most frequently cited terms related to the conceptual area in the languages and cultures of interest. This team then back-translates the items. Although an issue with traditional back-translation is the neglect of emic concepts in the target language, this process reduces this risk through the emphasis on commonalities between multiple target languages. After the back-translation process, the research team may identify the etic and emic conceptual dimensions. Items that appear frequently in different language groups signify a lower degree of emicity. Although concepts may appear in these multiple target cultures, Usunier points out that the facets of these concepts can be very different across cultures. Generating a cross-cultural inventory of a concept's facets can help researchers identify which facets are emphasized in differing cultures. Researchers should be careful in identifying the different facets of the concepts identified as largely etic. For racial and ethnic minority clinical research, the decentering process would involve the dominant European American culture and the culture of the specific racial and ethnic minority group under study (e.g., Latino/as or Asian Americans).

The final dimension of measurement equivalence is concerned with functional equivalence. If a construct from one culture serves a different or additional function in another culture, then functional equivalence may pose a problem in cross-cultural

studies. Functional equivalence in cross-cultural studies is also related to Cronbach and Meehl's (1955) concept of a nomological network of relations supporting construct validity. A measure is functionally equivalent if both the nature and pattern of relationships between the target measure and various constructs in the nomological network are similar across cultures. For example, nudity may be strongly associated with embarrassment in Culture A but not Culture B due to the lack of functional equivalence. Similarly, marriage has different functions in Western cultures, where romantic attraction is the primary foundation of such unions, whereas certain Muslim cultures allow for polygamy as long as the second, third, or fourth wife of the Muslim man converts to Islam in the marriage.

The assessment of functional equivalence of measures can occur via cross-cultural criterion-related and meta-analyses of effect sizes of those studies across culture groups, and as a program of research in itself. Evidence for functional equivalence can be demonstrated if cross-cultural criterion-related validity can be found. In other words, a target construct should be related to a theoretically relevant set of criterion variables across cultures. For example, major depression should be related to suicidal ideation across cultures. When a measure of major depression is related to suicidal ideation in one culture but not another, there should be concern about the functional equivalence of that measure. Similar to criterion-related studies, evidence from studies of concurrent validity of the target measure can also serve as evidence of functional equivalence. For example, the relationship between a new measure of depression and the Beck Depression Inventory should be the same across different cultures. When the relationship is not consistent across cultures, then questions arise as to the functional equivalence of the new measure.

Another source of information about functional equivalence of measures is meta-analytic studies of effect sizes across cultures. Cultural variations in effect sizes between target variables would raise concern about the functional equivalence of the measures for certain cultures. For example, Diener, Oishi, and Lucas (2003) have noted that ratings of subjective well-being are more influenced by self-serving biases for individuals from Western countries, whereas subjective well-being is more influenced by self-critical tendencies among East Asians. It appears that East Asians view measures involving self-evaluations as opportunities for self-improvement via criticism, whereas Westerners view such measures as opportunities for self-promotion and self-enhancement. More studies are needed to evaluate the extent to which lack of functional equivalence of constructs measured across racial and ethnic groups may be negatively affecting the development of theoretical models of psychopathology and psychotherapy.

Race and Ethnicity as Demographic Versus Psychological Variables

There have been two parallel approaches to the study of culture in psychology in the United States, namely the international and domestic approaches to multiculturalism (Leong et al., 2010). The former, labeled cross-cultural psychology, has its origins in anthropology and social psychology and is mainly concerned with comparing cultural differences between two or more countries (e.g., see articles in the *Journal of Cross-Cultural Psychology*). The latter, racial and ethnic minority psychology, has its origins in sociology and political science and is mainly concerned with racial and ethnic minority status and the associated disadvantages under the rubric of educational and income inequality or, more recently, health disparities. Given its historical origins in sociology, it is not accidental that racial and ethnic minority psychology has tended to adopt the paradigms of sociology and, therefore, investigates race and ethnicity as a social status variable with assumed disadvantages created by prejudice, discrimination, and racism. By treating race and ethnicity as a demographic variable, much of the past research in racial and ethnic minority psychology has generated findings that suffer from the "black box problem."

One conceptualization of the "black box problem" in psychological research criticizes the use of broad social categories (such as race and ethnicity) to represent homogeneous groups that do not warrant further decomposition. These groups are implicitly assumed to represent reality and the operationalization of the group variable is assumed to be sufficient as an explanatory construct. For example, the strong correlation between being Black and being a Democrat ignores the possibility that it may be social class that moderates the relationship between race and political party affiliation.

Within the field of psychology, there has been an ongoing controversy regarding the use of race and ethnicity as a demographic and not a psychological variable. While having its origins in biology and genetics, some have argued that race and ethnicity is currently a political and social construction with

complex ramifications. According to these scholars, using race and ethnicity as a demographic variable to serve as a proxy for such biological and genetic grouping of humans will result in overgeneralizations at best and gross misrepresentations at worst. Psychologists have therefore varied in their level of critique in using race and ethnicity as predictor variables.

For example, Matthews (1989) argues that associations between sociodemographic variables and health outcomes often covary with intervening psychological and behavioral factors that are not directly measured in health research. Whereas sociodemographic variables may serve as important marker variables, it is equally important not to conflate them with underlying causal mechanisms. For example, while rare, some men have been diagnosed with breast cancer, and it is a mistake to view breast cancer solely as a gender disease affecting women. In other words, we need to study the causal mechanisms underlying the metastatic process in the breast that tends to affect mainly women but also some men. Hence, we need to counter the tendency to overgeneralize from a correlation (e.g., between race/ethnicity and a particular mental health problem) to a causal inference. Demographic variables such as race, ethnicity, and gender can serve as a quick indicator of risk factors, but we need to differentiate between such social indicators versus psychological determinants of psychopathology.

Similarly, in reviewing research in cross-cultural psychology, Betancourt and Lopez (1993) found that studies have relied on presumed cultural and social features of a particular racial group to explain differences in racial groupings, a demographic variable, despite there being conceptual confusion in defining these features. They noted that "cross-cultural researchers who study cultural differences frequently fail to identify the specific aspects of culture and related variables that are thought to influence behavior ..... we learn that cultural group, race, or ethnicity may be related to a given psychological phenomenon, but we learn little about the specific elements of these group variables that contribute to the proposed relationship" (Betancourt & Lopez, 1993, p. 629). To resolve this problem, they proposed that both mainstream and cross-cultural investigators identify and measure directly the cultural element in a particular group of interest that is hypothesized to influence behavior. They noted that when culture (or race and ethnicity) is defined in terms of these psychologically relevant elements (e.g., values, beliefs, and acculturation), it becomes more amenable to measurement and "the relationship between these cultural elements to psychological phenomena can [therefore] be directly assessed" (p. 630).

Taking a more extreme position, Beutler, Brown, Crothers, Booker, and Seabrook (1996) suggest that using demographic variables like race, age, and sex without investigating the underlying psychological constructs of these demographic variables may result in conflicting or confusing findings. They delineate the conceptual limitations of demographic labels and question the scientific validity of using such grouping variables as race and ethnicity. Noting that racial self-referents are highly variable and arbitrary, they go on to observe that race as a biological construct is illusory. Despite the more extreme language, Beutler and colleagues (1996) hold a similar position to that of Matthews (1989) and Betancourt and Lopez (1993). They proposed that "psychosocial researchers and editors adopt a consistent definition of these terms and that research include an effort to identify the underlying concepts that the investigators assume to be reflected in these distinctions whenever these labels are used to report research findings" (Beutler et al., 1996, p. 892). A particularly valuable recommendation from their study is to treat race, ethnicity, and gender as psychological constructs and subject them to the same expectations of evaluating and demonstrating the construct validity of these variables in research.

More recently, in a debate regarding research on diversity and leadership published in the *American Psychologist*, Klein and Wang (2010) questioned the value of the review articles in terms of their primary focus on race and ethnicity as "surface-level diversity" as opposed to the "deep-level diversity" approach advancing in organizational psychology. In essence, their commentary is echoing the observations and recommendations of the earlier authors reviewed in this section with the distinction that we need to move beyond the use of surface-level diversity conceptualization, such as using race and ethnicity as a demographic variable without further delineating underlying psychological processes and mechanism. Deep-level diversity, on the other hand, seeks to "identify" and "unpack" these psychological processes and mechanisms. Cross-cultural clinical research also needs to move beyond using race and ethnicity as a demographic variable and begin to undertake deep-level diversity research that delves into the psychological processes and mechanisms that may be moderating or mediating the relationship between race/ethnicity and psychopathology.

From Confounds to Intersectionality

Another methodological challenge to cross-cultural clinical research is the role of confounds in research designs. A confound is defined as any variable that is not included in a research study but still has an effect on the research results. In their review of the psychotherapy literature, Zane and colleagues (2004) pointed to the neglect to control for variables that may be considered confounds with ethnicity or culture. In particular, they identified variables such as socioeconomic status, education level, and living environment that may have significant confounding effects with ethnicity or culture. They argued that failure to directly assess these variables may produce findings of ethnic or cultural differences with questionable internal validity. Furthermore, these variables may be significantly correlated with treatment outcomes and it would be useful to covary them to increase design sensitivity (Zane et al., 2004).

Following upon Zane and colleagues' (2004) critique of the neglect of these confounds in cross-cultural research, one solution would be to launch research programs that systematically examine the interaction between race/ethnicity and socioeconomic status or between race/ethnicity and gender. Indeed, such a program of research in psychology has arisen under the concept of intersectionality, which is concerned with "analytic approaches that simultaneously consider the meaning and consequences of multiple categories of identity, difference and disadvantage" (Cole, 2009, p. 170).

Prior to the call for intersectionality research, Cronbach (1957) had introduced the Attribute-Treatment-Interaction model for education and instruction. Cronbach had challenged the field to find "for each individual the treatment to which he can most easily adapt" with the observation that we can expect "some attributes of person to have strong interactions with treatment variables" (p. 681). Subsequent Attribute-Treatment-Interaction research found that some students with low ability performed better in highly structured treatments, whereas similar treatments hindered those with high abilities and preferences for less structured treatments. In a follow-up article, Cronbach (1975) emphasized the important relationship between cognitive aptitudes and treatment interactions and surmised that the inconsistency in his findings came from unidentified interactions.

Nevertheless, Cronbach's Attribute-Treatment-Interaction research set the stage for the learning orientation (i.e., learning styles) paradigm in educational psychology. This paradigm seeks to understand the structure and nature of the complex relationships between learning orientations and educational interventions (treatments). The delineation and exploration of this set of personal attributes that may interact with educational interventions served as the foundation for clinical research that sought to identify various client and therapist variables that interacted and influenced therapeutic outcomes. The Attribute-Treatment-Interaction paradigm therefore encouraged the examination of various personal attributes that may interact with treatment, which in turn served as the foundation for the concept of intersectionality. Instead of a single attribute such as cognitive ability, Attribute-Treatment-Interaction research spawned interest in multiple attributes that served as the stimulus for considering the simultaneous effects of several of these major attributes, such as gender, socioeconomic status, and religiosity.

In a recent review of intersectionality in psychology research, Cole (2009) reviewed the history of how feminist and critical race theory gave rise to this new approach that involved examining the conjoint effects of race, gender, and socioeconomic status. Cole cited a dearth of studies in the PsycInfo database that involve studies of two or more of these identity variables. The author continued to offer an approach to study intersectionality in psychological research organized around three questions: (1) "Who is included within this category?"; (2) "What role does inequality play?"; and (3) "Where are there similarities?" The first question encourages researchers to attend to the existing diversity within social categories and to examine the interdependency of these categories. The second question points out the existing hierarchies of power and privilege and challenges them. The last question attempts to find the similarities among the various categories that may appear very different at first sight. The goal of these three questions is to build upon each other and provide a rich context to advance our understanding of intersectionality.

Psychological research has tended to evolve from initial simple models to more complex ones in order to accurately capture and represent reality. Similar to the evolution of the DSM from a single clinical diagnosis to a multi-axial system of classification, Cronbach's (1957, 1975) Attribute-Treatment-Interaction model has encouraged and supported the recent attention to intersectionality in psychological research by challenging the "myth of client uniformity" (i.e., that all clients are the same and one treatment can work for all in every circumstance).

In the same way, the methodological challenge for cross-cultural clinical research is to adopt the Attribute-Treatment-Interaction paradigm in all of its complexity and examine the intersectionality of culture, race, ethnicity, gender, and social class in psychopathology and psychotherapy. As psychotherapy research evolved, the driving research question began to recognize the importance of intersectionality and was framed as "what works for whom, when and under what conditions?" In the same way, cross-cultural clinical research will also need to embrace an intersectionality or interactional perspective.

Differential Research Infrastructure

A final methodological challenge involves the problem of differential research infrastructure (Leong & Kalibatseva, 2010). Different subfields of psychology and psychiatry progress at different rates, and some are more advanced than others. For example, there have been considerably more studies of psychiatric epidemiology of White European American samples (as evidenced by the Epidemiological Catchment Area studies and the National Comorbidity Survey 1990–1992) than of African Americans, Latino/as, Asian Americans, Native Americans, and other minority groups. Despite oversampling of racial and ethnic minorities in the latest national epidemiological surveys, our knowledge base on psychiatric epidemiology for White European Americans is simply more advanced and more developed than for racial and ethnic minorities. In fact, the first National Latino and Asian American Study, a national epidemiological household study of Latino/as and Asian Americans in the United States, was completed in 2003. While recent clinical epidemiology studies have attempted to address the omission or underrepresentation of racial and ethnic minorities in previous samples, there is still a pressing need to conduct more nationally representative studies with these populations.

The problem of differential research infrastructure and the associated developmental lag is often overlooked or ignored by funding agencies, review boards, and even scientists themselves. Leaving aside the issue of the underlying causes for this differential research infrastructure across subfields, one factor has to do with the politics of numbers. Even a cursory review finds that some subfields do not receive attention or investment until a critical mass or critical number of agents and players are involved—sometimes referred to as the "tipping point." For example, in the field of human resource management, the *Workforce 2000* report (Johnston & Packer, 1987) from the Hudson Institute highlighted the impending demographic shifts in our country and alerted business leaders that we would be facing a significantly diverse workforce. This report led to increased attention to cultural diversity issues and initiatives in many organizations. Another example is the Supplement to the Surgeon General's Report on Mental Health, Culture, Race and Ethnicity in Mental Health (2001), which noted the significant ethnic minority mental health disparities and the critical knowledge gaps in those subfields. This and similar reports served as the impetus for increasing attention to health disparities, which eventually coalesced into a national priority.

The developmental lags across subfields can be seen as related to the differential research infrastructure (Leong & Kalibatseva, 2010). Specifically, subfields at earlier stages of development tend to have fewer investigators, journals, and grant-funded research projects, and less of an empirical base. The progress of any area of scientific inquiry will be proportional to the financial and human investments. By extension, just as health disparities research is concerned with correcting the differential (poorer) treatment and outcomes for different groups, we need to also be concerned with the research disparities created by the current differential research infrastructure across subfields where certain critical areas of cross-cultural clinical research suffer from a scarcity of researchers.

Leong and Kalibatseva (2010) provided an index of the differential research infrastructure and the associated research disparities by noting that for the year 1998–1999, there were 2,103 European Americans enrolled in Ph.D. psychology programs in comparison to 187 African Americans, 137 Hispanic Americans, and 217 Asian Americans. Conversely, for the year 2000, 2,601 European Americans received doctorates from graduate departments of psychology compared to 193 African Americans, 194 Hispanic Americans, and 149 Asian Americans. We cannot assume that all psychological scientists will conduct research related only to their own racial and ethnic groups, but one can get a sense of the differential research infrastructure related to mainstream versus racial and ethnic minority psychology. Similarly, in conducting a search in the PsycInfo database, Leong and Kalibatseva (2010) found 266,797 entries for the word "depression," of which only 7,983 studies concerned depression among African Americans and 1,948 studies concerned depression among Asian

Americans. As a final illustration of the problem, Leong and Kalibatseva (2011) noted that *American Psychologist* was established in 1946, whereas the *Journal of Black Psychology* was established in 1974 and the *Asian American Journal of Psychology* was established in 2010.

In summary, we have chosen to highlight the critical problems of differential research infrastructure and the shortage of human capital that serve as barriers to the advancement of racial and ethnic minority clinical research. The contextual factors contributing to these research disparities need to be included as important elements in our national plan to advance our understanding and improvement of the mental health of racial and ethnic minorities. Attending to racial and ethnic disparities in mental health services without attending to the underlying differential research infrastructure is like trying to improve the academic performance of the students in educational institutions without improving the quality and training of teachers.

References

Alloy, L. B., Abramson, L. Y., Raniere, D., & Dyllere, I. M. (2003). Research methods in adult psychopathology. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed.) (pp. 466–498). New York: John Wiley & Sons.

American Psychiatric Association. (1952). *Diagnostic and statistical manual: Mental disorders*. Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist*, *58*, 377–402.

American Psychological Association. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*, 271–285.

Atkinson, D. R., Bui, U., & Mori, S. (2001). Multiculturally sensitive empirically supported Treatments—An oxymoron? In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (2nd ed., pp. 542–574). Thousand Oaks, CA: Sage.

Berry, J. W. (1980). Acculturation as varieties of adaptation. In A. M. Padilla (Ed.), *Acculturation: Theory, models and some new findings* (pp. 9–25). Boulder, CO: Westview Press.

Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist*, *48*, 629–637.

Beutler, L. E., Brown, M. T., Crothers, L., Booker, K., & Seabrook, M. K. (1996). The dilemma of factitious demographic distinctions in psychological research. *Journal of Consulting and Clinical Psychology*, *64*, 892–902.

BigFoot, D. S., & Schmidt, S. (2010). Honoring children, mending the circle: Cultural adaptation of trauma-focused cognitive-behavioral therapy for American Indian and Alaska Native children. Journal of Clinical Psychology, 66, 847–856.

Brislin, R. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216.

Brislin, R. (1980). Translation and content analysis of oral and written materials. In H. Triandis & J. Berry (Eds.), *Handbook of cross-cultural psychology: Volume 2, Methodology* (pp. 389–444). Boston: Allyn and Bacon.

Brislin, R. (2000). *Understanding culture's influence on behavior* (2nd ed.). Fort Worth, TX: Harcourt.

Bucardo, J. A., Patterson, T. L., & Jeste, D. V. (2008). Cultural formulation with attention to language and cultural dynamics in a Mexican psychiatric patient treated in San Diego, California. *Culture, Medicine, & Psychiatry*, *32*, 102–121.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology*, *3*, 94–105.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Chambless, D. L., Sanderson, W. C., Shoham, V., Johnson, S. B., Pope, K. S., Crits-Christoph, P., et al. (1996). An update on empirically validated therapies. *Clinical Psychologist*, *49*, 5–18.

Cheung, F., & Lin, K.-M . (1997). Neurasthenia, depression and somatoform disorder in a Chinese-Vietnamese woman immigrant. *Culture, Medicine, & Psychiatry*, *21*, 247–258.

Cochrane, A. L. (1979). 1931–1971: A critical review with particular reference to the medical profession. In Please provide editor name(s). *Medicines for the year 2000* (pp. 1–11). London: Office of Health Economics.

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*, 170–180.

Constantine, M. G., & Sue, D. W. (2006). Factors contributing to optimal human functioning in people of color in the United States. *Counseling Psychologist*, *34*, 228–244.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cronbach, L. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Diener, E., Oishi, S., & Lucas, R. E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology*, *54*, 403–425.

Ecklund, K., & Johnson, W. B. (2007). The impact of a culture-sensitive intake assessment on treatment of a depressed biracial child. *Clinical Case Studies*, *6*, 468–482.

Fernandez, K., Boccaccini, M. T., & Noland, R. M. (2008). Detecting over—and underreporting of psychopathology with the Spanish-language Personality Assessment Inventory: Findings from a simulation study with bilingual speakers. *Psychological Assessment*, *20*, 189–194.

Garfield, S. (1984). Methodological problems in clinical diagnosis. In H. E. Adams & P. B. Sutker (Eds.), *Comprehensive handbook of psychopathology* (pp. 27–44). New York: Plenum.

Gelso, C. J. (1979). Research in counseling: Methodological and professional issues. *Counseling Psychologist*, *8*, 7–35.

Gelso, C. J. (1992). *Counseling psychology*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Griner, D., & Smith, T. B. (2006). Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy: Theory, Research, Practice, Training*, *43*, 531–548.

Hall, G. C. N. (2001). Psychotherapy research with ethnic minorities: Empirical, ethical, and conceptual issues. *Journal of Consulting and Clinical Psychology*, *69*, 502–510.

Hall, G. C. N., & Eap, S. (2007). Empirically-supported therapies for Asian Americans. In F. T. L. Leong, A. Inman, A. Ebreo, L. Yang, L. Kinoshita, & M. Fu (Eds.), *Handbook of Asian American psychology* (2nd ed., pp. 449–467). Thousand Oaks, CA: Sage.

Hall, G. C. N., Hong, J. J., Zane, N. W. S., & Meyer, O. L. (2011). Culturally competent treatment for Asian Americans: The relevance of mindfulness and acceptance-based psychotherapies. *Clinical Psychology Science and Practice, 18,* 215–231.

Hall, G. C. N., & Maramba, G. G. (2001). In search of cultural diversity: Recent literature in cross-cultural and ethnic minority psychology. *Cultural Diversity & Ethnic Minority Psychology*, *7*, 12–26.

Henriksen, R. C., & Paladino, D. A. (2009). *Counseling multiple heritage individuals, couples, and families*. Alexandria, VA: American Counseling Association.

Hofferth, S. L. (2005). Secondary data analysis in family research. *Journal of Marriage and Family*, *67*, 891–907.

Huey, S. J., & Polo, A. J. (2008). Evidence-based psychosocial treatments for ethnic minority youth. *Journal of Clinical and Adolescent Psychology*, *37*, 262–301.

Hwang, W.C . (2006). The psychotherapy adaptation and modification framework: Application to Asian Americans. *American Psychologist*, *61*, 702–715.

Johnston, W. B., & Packer, A. E., (1987). *Workforce 2000. Work and workers for the 21st century*. Indianapolis, IN: The Hudson Institute, Inc.

Kanter, J. W., Santiago-Rivera, A. L., Rusch, L. C., Busch, A. M., & West, P. (2010). Initial outcomes of a culturally adapted behavioral activation for Latinas diagnosed with depression at a community clinic. *Behavior Modification*, *34*, 120–144.

Kazdin, A. E. (2008). 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.

Klein, K. M., & Wang, M. (2010). Deep-level diversity and leadership. *American Psychologist*, *65*(9), 932–934.

Kluckhohn, C., & Murray, H. A. (1950). Personality formation: The determinants. In C. Kluckhohn & H. A. Murray (Eds.), *Personality in nature, society, and culture* (pp. 35–48). New York: Knopf.

Kohn, L. P., Oden, T., Munoz, R. F., Robinson, A., & Leavitt, D. (2002). Adapted cognitive behavioral group therapy for depressed low-income African American women. *Community Mental Health Journal*, *38*, 497–504.

Lau, A. (2006). Making the case for selective and directed cultural adaptations of evidence-based treatments: Examples from parent training.*Clinical Psychology:Science and Practice*, *13*, 295–310.

Leong, F. T. L. (1996). Towards an integrative model for cross-cultural counseling and psychotherapy. *Applied and Preventive Psychology*, *5*, 189–209.

Leong, F. T. L. (2007). Cultural accommodation as method and metaphor. *American Psychologist*, *62*(8), 916–927.

Leong, F. T. L., & Kalibatseva, Z. (2010). Comparative effectiveness research on Asian American mental health: Review and recommendations. *AAPI Nexus*, *8*(2), 21–38.

Leong, F. T. L., & Kalibatseva, Z. (2011). Effective psychotherapy for Asian Americans: From cultural accommodation to cultural congruence. *Clinical Psychology: Science and Practice*, *18*, 242–245.

Leong, F. T. L., & Lee, S. H. (2006). A cultural accommodation model for cross-cultural psychotherapy: Illustrated with the case of Asian Americans. *Psychotherapy: Theory, Research, Practice, Training*, *43*(4), 410–423.

Leong, F. T. L., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology*, *16*(4), 590–597.

Leong, F. T. L., & Lopez, S. (2006). Guest editors' introduction. *Psychotherapy: Theory, Research, Practice, Training*, *43*(4), 378–379.

Leong, F. T. L., & Serafica, F. (2001). Cross-cultural perspectives on Super's career development theory: Career maturity and cultural accommodation. In F. T. L. Leong & A. Barak (Eds.), *Contemporary models in vocational psychology: A volume in honor of Samuel H. Osipow* (pp. 167–205). Mahwah, NJ: Erlbaum.

Leong, F. T. L., & Tang, M. (2002). A Cultural Accommodation approach to career assessment with Asian Americans. In K. Kurasaki, S. Sue, & S. Okazaki (Eds.), *Asian American mental health: Assessment, theories and methods* (pp. 265–281). The Netherlands: Kluwer Academic Publishers.

Leong, F. T. L., Wagner, N. S., & Tata, S. (1995). Racial and ethnic variations in help-seeking attitudes. In J. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (pp. 415–438). Thousand Oaks, CA: Sage.

Li, L. C., Kim, B. S. K, & O'Brien, K. M. (2007). An analogue study of the effects of Asian cultural values and counselor multicultural competence on counseling process. *Psychotherapy: Theory, Research, Practice, Training*, *44*, 90–95.

Lopez, S. R., & Guarnaccia, P. J. J. (2000). Cultural psychopathology: Uncovering the social world of mental illness. *Annual Review of Psychology*, *51*, 571–598.

Matarazzo, J. D. (1978). The interview: Its reliability and validity in psychiatric diagnosis. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A handbook* (pp. 47–96). New York: Plenum.

Matthews, K. A. (1989). Are sociodemographic variables markers for psychological determinants of health? *Health Psychology*, *8*, 641–648.

Mezzich, J. E., Kirmayer, L. J., Kleinman, A., Fabrega, H., Parron, D., Good, B., et al. (1999). The place of culture in DSM-IV. *Journal of Nervous and Mental Disease*, *187*, 457–464.

Miranda, J., Bernal, G., Lau, A., Kohn, L., Hwang, W.-C., & LaFromboise, T. (2005). State of the science on psychosocial interventions for ethnic minorities. *Annual Reviews in Clinical Psychology*, *1*, 113–142.

Miranda, J., Nakamura, R., & Bernal, G. (2003). Including ethnic minorities in mental health intervention research: A practical approach to a long-standing problem. *Culture, Medicine and Psychiatry*, *27*, 467–486.

Muñoz, R. F., & Mendelson, T. (2005). Toward evidence-based interventions for diverse populations: The San Francisco General Hospital prevention and treatment manuals. *Journal of Consulting and Clinical Psychology*, *73*, 790–799.

Pan, D., Huey, S. J., & Hernandez, D. (2011). Culturally adapted versus standard exposure treatment for phobic Asian Americans: Treatment efficacy, moderators, and predictors. *Cultural Diversity and Ethnic Minority Psychology*, *17*, 11–22.

Pina, A. A., Villalta, I. K., & Zerr, A. A. (2009). Exposure-based cognitive behavioral treatment of anxiety in youth: An emerging culturally-prescriptive framework. *Behavioral Psychology*, *17*, 111–135.

Rosenhan, D. L. (1973). On being sane in insane places. *Science*, *179*, 250–258.

Rossello, J., Bernal, G., & Rivera, C. (2008). Randomized trial of CBT and IPT in individual and group format for depression in Puerto Rican adolescents. *Cultural Diversity and Ethnic Minority Psychology*, *14*, 234–245.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210–222.

Shore, J. H., & Manson, S. M. (2004). The American Indian veteran and posttraumatic stress disorder: A telehealth assessment and formulation. *Culture, Medicine, & Psychiatry*, *28*, 231–243.

Smith, T. B., Domenech Rodriguez, M. & Bernal, G. (2011). Culture. *Journal of Clinical Psychology: In Session*, *67*, 166–175.

Sue, S., Kurasaki, K. S., & Srinivasan, S. (1999). Ethnicity, gender, and cross-cultural issues in clinical research. In P. C. Kendall, J. N. Butcher, & G. N. Hombeck (Eds.), *Handbook of research methods in clinical psychology* (2nd ed., pp. 54–71). New York: Wiley.

Sue, S., Fujino, D. C., Hu, L., Takeuchi, D., & Zane, N. S. W. (1991). Community mental health services for ethnic minority groups: A test of the cultural responsiveness hypothesis. *Journal of Consulting and Clinical Psychology*, *59*, 533–540.

Trimble, J. E. (2005). An inquiry into the measurement of ethnic and racial identity. In R. T. Carter (Ed.), *Handbook of racial-cultural psychology and counseling, Vol. 1: Theory and research* (pp. 320–359). Hoboken, NJ: John Wiley & Sons Inc.

Trzesniewski, K. H., Donnellan, M. B., & Lucas, R. E. (Eds.) (2010). *Secondary data analysis: A guide for psychologists*. Washington, DC: APA.

U.S. Census Bureau (2011). *Projected population by single year of age, sex, race, and Hispanic origin for the United States: July 1, 2000 to July 1, 2050*. Retrieved from http://www.census.gov/population/www/projections/downloadablefiles.html

U.S. Department of Health and Human Services (1999). *Mental health: A report of the surgeon general*. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.

U.S. Department of Health and Human Services. (2001). *Mental health: Culture, race, and ethnicity. A supplement to Mental health: A report of the surgeon general*. Rockville, MD: U.S. Department of Health and Human Services. Retrieved from http://www.surgeongeneral.gov/library/mentalhealth/cre/sma-01–3613.pdf.

Usunier, J. C. (1998). *International & cross-cultural management research*. London: Sage.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Whaley, A. L., & Davis, K. E. (2007). Cultural competence and evidence-based practice in mental health services: A complimentary perspective. *American Psychologist*, *62*, 563–574.

Zane, N., Hall, G. C. N., Sue, S., Young, K., & Nunez, J. (2004). Research on psychotherapy with culturally diverse populations. In M. J. Lambert (Ed.), *Psychotherapy and behavior change* (5th ed., pp. 767–804). New York: Wiley.

This page intentionally left blank

PART 5

Conclusion

This page intentionally left blank

CHAPTER

22 Decades Not Days: The Research Enterprise in Clinical Psychology

Philip C. Kendall *and* Jonathan S. Comer

Abstract

Acknowledging the nascent stage of the collaboration between science and clinical psychology, this chapter discusses features of the preferred approach to guide the accumulation of knowledge through careful implementation and evaluation. The collaboration of science and practice in clinical psychology has produced meaningful, reliable, and replicated findings, and illustrations (e.g., exposure for anxiety, cognitive-behavioral therapy for depression and panic) are identified. Speculations regarding the future of our field and a model for optimal science–practice collaborations in clinical psychology are offered.

Key Words: Clinical science, empirically supported treatment, scientist-practitioner

From the earliest days of our field, scholars have asked the important questions about the causes, nature, and amelioration of mental health problems, and a multitude of "answers" have been offered in as much time. However, it is only recently, relatively speaking, that the scientific method and empiricism has been applied to the field of clinical psychology. And thus, whereas many sciences have been meaningfully progressing and providing valid and reliable answers for centuries (e.g., biology, chemistry, physics), we remain at a relatively early stage in the science of clinical psychology, with the majority of work ahead of us. Needless to say, we have quite a bit of catching up to do.

Importantly, despite our field's "late start" in the application of the scientific method, rigorous and relevant science cannot be rushed or streamlined. Like pet-producing "puppy mills," where dog breeding is reduced to the birthing of large quantities of animals with limited regard to the vicissitudes of inbreeding or routine care, research laboratories that produce more in quantity but less in quality are not of great merit. A research mill might produce a long list of publications, some of which likely appear in vanity journals or pay-to-publish outlets, but it typically contributes little to the advancement of a field. The rapid generation of multiple separate papers from limited datasets and the absence of a coherent integration of collected findings lack keys to advancing knowledge. Research mills are far from the goal of this collection of writings on research strategies.

Magazine covers, blog entries, and brief television and radio news broadcasts can bring instant publicity to a finding, a degree of fame to the investigator, and a momentary splash of excitement for the findings. Studies with more splash than substance are hoisted by newspersons for the instant gratification of a boost in ratings. Splash is not typically associated with one of the most important features of good research—being replicated. Much of the news that qualifies as splash fades rapidly over time, and like food fads, clothing trends, and the whims of teens, splash lacks longevity.

A scientific enterprise connotes a long and persistent undertaking, an undertaking that consists of multiple smaller activities. A scientific enterprise in clinical psychology, when adopting the most

rigorous strategies across a diversity of research methods, carries the greatest promise for advancement. Clinical psychology has and will continue to advance as those asking and answering the important questions of the day contribute the decade(s) of work that makes up an enterprise.

How best should researchers in clinical psychology navigate this scientific enterprise in a manner that maximizes both rigor and relevance? Our collection of expert psychology methodologists has provided the ultimate in timely and readable advice for future research. The methodological niceties have been described and the working processes have been detailed. Application of the advice and procedures will no doubt enhance our discipline.

Importantly, however, those looking for the "correct" research strategy with which to address all questions are misguided. In the introduction of this Handbook, we underscored that: just as with any travel directions, where many acceptable ways to get to the same destination may exist (e.g., the quick way, the scenic way, the cheap way), for each empirical question there are many research strategies that can be used to reveal meaningful information, each with strengths and limitations. When conducting research, it is incumbent upon the investigator to explicitly know why he or she is taking a particular route, to be familiar with the tradeoffs inherent in taking such a route, and to travel that route correctly. Collectively, the works in this volume detail a portfolio of modern research strategies for the science of clinical psychology—a set of alternative and complementary "directions," so to speak, for advancing our field from where we are now to where we need to get.

Space limitations prevent an exhaustive list of the many meaningful, reliable, and replicated findings our field has already generated through diverse and sustained efforts across decades in the grand research enterprise. Rather, a few illustrative examples will suffice:

1. *Exposure for Anxiety*: We know that simply listening and perhaps even unwittingly accommodating to anxious distress is not as good as facing anxiety in an exposure task. Clinging to arcane notions of "support" in the absence of exposure might even serve to unwittingly maintain anxiety (e.g., Abramowitz et al., 2011; Foa, 2011; Foa, Rothbaum, & Furr, 2003; Hazlett-Stevens & Craske, 2009; Heimberg, 2002; Kendall et al., 2005, 2008; Silverman et al., 2008; Walkup et al., 2008). Long-term follow-up evaluations with adolescents and young adults treated previously as anxious children show the gains associated with exposure-based treatments and their endurance (Kendall & Southam-Gerow, 1996; Kendall et al., 2004; Saavedra, Silverman, et al., 2010).

2. *Cognitive-Behavioral Therapy (CBT) for Depression*: We know so much more now than decades ago about the nature and neurobiology of cognitive processing in depression (Abramson, Alloy et al., 2002; Clark & Beck, 2010) and about the features of effective cognitive therapy and CBT for depression (for example, see Jacobson et al., 2001), and we now know that CBT/cognitive therapy, when used as a monotherapy or in conjunction with supported pharmacology, can have beneficial effects on episodic depression that endure long after treatment is discontinued (DeRubeis et al., 2008; Dobson et al., 2008; Hollon et al., 2006).

3. *Behavioral Parent Training for Early Disruptive Behavior Problems*: Disruptive behavior disorders—characterized by problems of conduct and oppositionality—emerge in early childhood, exhibit considerable stability, and are associated with profound disability. Effective early intervention is critical. After decades of focused research, there is now support for psychological interventions targeting early child behavior problems indirectly by reshaping parent practices, with the goals of increasing in-home predictability, consistency, and follow-through and promoting effective discipline (Comer et al., 2013; Eyberg et al., 2008; Gleason et al., 2007). These treatments help families disrupt coercive cycles by training parents to increase positive feedback for appropriate behaviors, ignore disruptive behaviors, and provide consistent time-outs for noncompliance. Efficacious parent management programs have shown enduring support across time, powerfully offsetting unfavorable trajectories toward truancy, substance use, and criminality in a great proportion of affected youth.

4. *CBT for Panic Disorder and Agoraphobia*: Once believed to be a puzzling and nonspecific form of free-floating anxiety, through experimental, clinical, biological challenge, and longitudinal research, we now understand that panic disorder is an acquired fear of certain bodily sensations (especially those elicited by autonomic arousal) and that agoraphobia is a pattern of avoidance in response to the anticipation of bodily sensations (Barlow, 1988; Clark, 1986; Craske & Barlow, 2001; Ehlers & Margraf, 1989). Building

on these findings, panic control treatments that incorporate cognitive restructuring targeting misappraisals of bodily sensations, as well as exposure to avoided situations and autonomic arousal (i.e., interoceptive exposures), have shown substantial advantage over comparison treatments containing a therapeutic alliance and positive patient expectancies (see Craske & Barlow, 2001). Across rigorous clinical trials, such CBTs compare quite favorably with leading psychotropic interventions and are even more durable over the long term (Barlow, Gorman, Shear, & Woods, 2000).

5. *Assessments that Predict*: The earliest assessments in clinical psychology regrettably often relied exclusively on subjective interpretation and ultimately proved unreliable and invalid. Psychological assessment in clinical psychology evolved to incorporate rationally generated items according to face validity, but it was not until the psychometric work of Hathaway and McKinley developing the Minnesota Multiphasic Personality Inventory (MMPI) during World War II that empirical demonstration of a test's ability to differentiate among predetermined groups became a routine concern in clinical psychology (Butcher, 2010; Graham, 2000). Since the inception of the MMPI, empirical methods to test construction and evaluation have become the norm in clinical psychology research, and the tremendous advances across the ensuing decades in the science of psychometrics and quality assessment—including more recent activities in item-response theory—have laid the very reliable and valid foundation upon which the field of clinical science now sits.

A Future Speculation

Past accomplishments are worthy recollections that buttress the notion that clinical research and practice can coexist and collaborate, and a round of applause is heard to document the widespread support for this coexistence. But what lies ahead? What might, in the next 30 years, be recalled as one of the great advances of the scientific enterprise in clinical psychology?

Although multiple potential examples come to mind, one overarching realm of scholarly pursuit may serve to be particularly fruitful in the coming years, given its simultaneous integration of nomothetic methods and an idiographic perspective on individual differences. This area is the elucidation of mediators and moderators of psychopathology course and treatment effects. For present purposes, let's consider potential moderators of treatment response.

As detailed in Chapter 15 of this Handbook, a moderator in the context of psychotherapy is a variable that delineates the conditions under which a given treatment is related to an outcome. Moderators identify *on whom* and *under what circumstances* which treatments have different effects (after Kiesler, 1971; Kraemer et al., 2002). A moderator is a variable that influences either the direction or the strength of a relationship between treatment and outcome. For example, if the results of a randomized controlled trial indicate that one of the studied treatments was more effective with women than with men, but this gender difference was not found in response to a comparison treatment, then gender would be a moderator of the association between treatments and outcome. Treatment moderators clarify for clinicians (and other consumers of research) which clients might be most responsive to which treatments and for which clients alternative treatment might be sought. Gender in this example is a straightforward example of a moderator, whereas comorbid condition, initial severity, life stress, a biological state, motivation, and other variables are somewhat more complex potential moderators. When a variable is associated broadly with outcome across treatment conditions, that variable is simply a *predictor* of outcome, but not a treatment moderator. Such predictor variables are less informative, as they indiscriminately predict whether someone will benefit from any treatment, rather than providing prescriptive information about which treatment may produce the optimal results under a given circumstance.

Given the tremendous advances over the past several decades in the development of a wide range of evidence-based treatments for various conditions, a scientific enterprise applying the full spectrum of research strategies outlined throughout the present Handbook to uncover moderators of treatment response and psychology course has the enormous bridging potential to provide an evidence-based roadmap, so to speak, for the application of psychological science to both public policy on a grand scale as well as individual practice and clinical decision making.

A Guiding Model

The scientist-practitioner model captures the clinician and clinical researcher as inseparable contributors to and consumers of advances in knowledge.

The scientist-practitioner model is widespread. Indeed, both scientists who evaluate their work and their theories with rigor, and practitioners who utilize a research-based understanding of human behavior in social contexts to aid people in resolving psychological dysfunctions and enhancing their lives, follow the scientist-practitioner model. The ideal is not intended to create professional role confusion but rather to foster service providers who evaluate their interventions scientifically and researchers who study applied questions and interpret their findings with an understanding of the richness and complexity of human experience (Kendall & Norton-Ford, 1982). For treatment outcome studies to be meaningful and relevant, they must reflect both a fit within the guidelines of science and an understanding of the subtleties of human experience and behavior change. Exceptionally controlled investigations that are distant from the realities of therapy may offer only limited conclusions. Uncontrolled studies of therapy fail to pinpoint the effects that can be accurately attributed to therapy and provide, at best, speculations. The scientist-practitioner develops a variety of methods for studying meaningful therapeutic interventions and outcomes in a scientific fashion. Indeed, one can find fairly widespread acceptance of the scientist-practitioner model and commitment to it when we consider the question, "What if we did not seek empirical evaluation of the effects of therapy?" (Beutler, 1998; Kendall, 1998). What process would replace it as we seek to advance our understanding of what treatments are effective for what types of client problems? If we didn't use empirically supported treatments, would we rely on each therapist's own views, which vary and which are, as are all our views, slanted by personal experience?

Although our relatively young field has amassed an impressive base of empirically based knowledge and applied know-how, many of the most important answers are ahead of us. And, although some answers may elude us, we advance as we tolerate reducing error variance and gradual progress. Such a scientific enterprise may lack some of the transient splash associated with creative speculations and subjective musings based on introspection alone, but in the long term it is the scientific method and a commitment to empiricism that meaningfully advances our field and truly improves the lives of patients.

A Potential Dedication

Thanks to you for your dedication to quality research in clinical psychology. The purpose of this book is to advance quality research. This book could be dedicated "to the napkins at local restaurants, where some of the most innovative research has been initiated, revised, and designed." We recognize that such a dedication may seem idiosyncratic, but for many readers it may ring true. It is often at casual discussions over a meal when researchers and clinicians most effectively think through the pluses and limitations of research designs and, in the absence of a pad of paper or a laptop, sketch out designs on unfolded napkins. This dedication also underscores our strongly held belief that the application of the most rigorous scientific strategies to the most clinically relevant questions is necessary but not sufficient for the highest caliber of research. Indeed, the missing ingredient for the best research is the passion of the investigator—the passion that could never wait for a proper piece of paper when the inspiration hits. It is this passion that can never be taught in a handbook such as this, and it is this passion that you, the reader, must bring.

Acknowledgments

Preparation of this work was facilitated by research grants (MH063747; MH086438) awarded to Philip C. Kendall, and (K23 MH090247) awarded to Jonathan Comer.

References

Abramowitz, J. S., Deacon, B. J., & Whiteside, S. P. H. (2011). *Exposure therapy for anxiety: Principles and practice*. New York: Guilford Press.

Abramson, L. Y., Alloy, L. B., Hogan, M. E., Whitehouse, W. G., Donovan, P., Rose, D. T., &…Raniere, D. (2002). Cognitive vulnerability to depression: Theory and evidence. In R. L. Leahy & E. Dowd (Eds.), *Clinical advances in cognitive psychotherapy: Theory and Application* (pp. 75–92). New York: Springer.

Barlow, D. H. (1988). *Anxiety and its disorders: The nature and treatment of anxiety and panic*. New York: Guilford Press.

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2000). Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *Journal of the American Medical Association*, *283*, 2529–2536.

Butcher, J. N. (2010). Personality assessment from the nineteenth to the early twenty-first century: Past achievements and contemporary challenges. *Annual Review of Clinical Psychology*, *6*, 1–20.

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy*, *24*, 461–470.

Clark, D. A., & Beck, A. T. (2010). Cognitive theory and therapy of anxiety and depression: Convergence with neurobiological findings. *Trends in Cognitive Sciences*, *14*(9), 418–424.

Comer, J.S., Chow, C., Chan, P., Cooper-Vince, C., & Wilson, L.A.S. (2013). Psychosocial treatment efficacy for disruptive behavior problems in young children: A meta-analytic

examination. *Journal of the American Academy of Child and Adolescent Psychiatry, 52*, 26–36.

Craske, M. G., & Barlow, D. H. (2001). Panic disorder and agoraphobia. In D. H. Barlow (Ed.), *Clinical handbook of psychological disorders: A step-by-step treatment manual* (3rd ed., pp. 1–59). New York: Guilford.

DeRubeis, R. J., Siegle, G. J., & Hollon, S. D. (2008). Cognitive therapy versus medication for depressions: Treatment outcomes and neural mechanisms. *Nature Reviews Neuroscience*, *9*(10), 788–796.

Dobson, K. S., Hollon, S. D., Dimidjian, S., Schmaling, K. B., Kohlenberg, R. J., Gallop, R. J., & … Jacobson, N. S. (2008). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. *Journal of Consulting and Clinical Psychology*, *76*, 468–477.

Ehlers, A., & Margraf, J. (1989). The psychophysiological model of panic attacks. In P. M. G. Emmelkamp (Ed.), *Anxiety disorders: Annual series of European research in behavior therapy* (Vol. 4, pp. 1–29). Amsterdam: Swets & Zeitlinger.

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology*, *37*, 215–237.

Foa, E. B. (2011). Prolonged exposure therapy: Past, present, and future. *Depression and Anxiety*, *28*, 1043–1047.

Foa, E. B., Rothbaum, B. O., & Furr, J. M. (2003). Augmenting exposure therapy with other CBT procedures. *Psychiatric Annals*, *33*, 47–53.

Gleason, M. M., Egger, H. L., Emslie, G. J., Greenhill, L. L., Kowatch, R. A., Lieberman, A. F., Luby, J. L., Owens, J., Scahill, L. D., Scheeringa, M. S., Stafford, B., Wise, B., & Zeanah, C. H. (2007). Psychopharmacological treatment for very young children: Contexts and guidelines. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*,1532–1572.

Graham, J. R. (2000). *MMPI-2: Assessing personality and psychopathology* (3rd ed.). New York: Oxford.

Hazlett-Stevens, H., & Craske, M. G. (2009). Live (in vivo) exposure. In W. T. O'Donohue & J. E. Fisher (Eds.), *General principles and empirically supported techniques of cognitive behavior therapy* (pp. 407–414). Hoboken, NJ: John Wiley & Sons Inc.

Heimberg, R. G. (2002). Cognitive-behavioral therapy for social anxiety disorder: Current status and future directions. *Biological Psychiatry*, *51*, 101–108.

Hollon, S. D., Stewart, M. O., & Strunk, D. (2006). Enduring effects for cognitive behavior therapy in the treatment of depression and anxiety. *Annual Review Of Psychology*, *57*, 285–315.

Jacobson, N. S., Martell, C. R., & Dimidjian, S. (2001). Behavioral activation treatment for depression: Returning to contextual roots. *Clinical Psychology: Science and Practice*, *8*(3), 255–270.

Kendall, P. C., & Norton-Ford, J. D. (1982). *Clinical psychology: Scientific and professional dimensions*. New York: Wiley.

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology*, *76*(2), 282–297.

Kendall, P. C., Robin, J. A., Hedtke, K. A., Suveg, C., Flannery-Schroeder, E., & Gosch, E. (2005). Considering CBT with anxious youth? Think exposures. *Cognitive and Behavioral Practice*, *12*, 136–150.

Kendall, P. C., Safford, S., Flannery-Schroeder, E., & Webb, A. (2004). Child anxiety treatment: outcomes in adolescence and impact on substance use and depression at 7.4-year follow-up. *Journal of Consulting and Clinical Psychology*, *72*(2), 276–287.

Kendall, P. C., & Southam-Gerow, M. A. (1996). Long-term follow-up of a cognitive-behavioral therapy for anxiety-disordered youth. *Journal of Consulting and Clinical Psychology*, *64*, 724–730.

Kiesler, D. J. (1971). Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (pp. 36–74). New York: Wiley

Kraemer, H., Wilson, G., Fairburn, C. G., & Agras, W. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877–883.

Saavedra, L. M., Silverman, W. K., Morgan-Lopez, A. A., & Kurtines, W. M. (2010). Cognitive behavioral treatment for childhood anxiety disorders: Long-term effects on anxiety and secondary disorders in young adulthood. *Journal of Child Psychology and Psychiatry*, *51*(8), 924–934.

Silverman, W. K., Pina, A. A., & Viswesvaran, C. (2008). Evidence-based psychosocial treatments for phobic and anxiety disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *37*, 105–130.

Walkup, J. T., Albano, A., Piacentini, J., Birmaher, B., Compton, S. N., Sherrill, J. T., &… Kendall, P. C. (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine*, *359*, 2753–2766.

This page intentionally left blank

INDEX

Note: Page references followed by "*f*" and "*t*" denote figures and tables, respectively.