APA Handbook of

Testing and Assessment in Psychology

APA Handbook of Testing and Assessment in Psychology, Vol. 2: Testing and Assessment in Clinical and Counseling Psychology, edited by K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez

APA Handbooks in Psychology

APA Handbook of Testing and Assessment in Psychology

VOLUME 2

Testing and Assessment in Clinical and Counseling Psychology

Kurt F. Geisinger, *Editor-in-Chief*

Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodriguez, *Associate Editors*

American Psychological Association • Washington, DC

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/pubs/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

American Psychological Association Staff
Gary R. VandenBos, PhD, *Publisher*
Julia Frank-McNeil, *Senior Director, APA Books*
Theodore J. Baroody, *Director, Reference, APA Books*
Lisa T. Corry, *Project Editor, APA Books*

Typeset in Berkeley by Cenveo Publisher Services, Columbia, MD

Printer: United Book Press, Baltimore, MD
Cover Designer: Naylor Design, Washington, DC

Library of Congress Cataloging-in-Publication Data

APA handbook of testing and assessment in psychology / Kurt F. Geisinger, editor-in-chief ; Bruce A. Bracken . . . [et al.], associate editors.
v. cm. — (APA handbooks in psychology)
Includes bibliographical references and index.
Contents: v. 1. Test theory and testing and assessment in industrial and organizational psychology — v. 2. Testing and assessment in clinical and counseling psychology — v. 3. Testing and assessment in school psychology and education.
ISBN 978-1-4338-1227-9 — ISBN 1-4338-1227-4
1. Psychological tests. 2. Psychometrics. 3. Educational tests and measurements. I. Geisinger, Kurt F., 1951- II. Bracken, Bruce A. III. American Psychological Association. IV. Title: Handbook of testing and assessment in psychology.
BF176.A63 2013
150.28′7—dc23
2012025015

British Library Cataloguing-in-Publication Data
A CIP record is available from the British Library.

Printed in the United States of America
First Edition

DOI: 10.1037/14048-000

Contents

Volume 2: Testing and Assessment in Clinical and Counseling Psychology

Editorial Board

PART I

GENERAL ISSUES IN TESTING AND ASSESSMENT IN PROFESSIONAL PSYCHOLOGY

CHAPTER 1

CLINICAL AND COUNSELING TESTING

Janet F. Carlson

Many clinical and counseling psychologists depend on tests to help them understand as fully as possible the clients with whom they work (Camara, Nathan, & Puente, 2000; Hood & Johnson, 2007; Masling, 1992; Naugle, 2009). A broad and comprehensive understanding of an individual supports decisions to be made by or regarding a client. Tests provide a means of sampling behavior, with results used to promote better decision making. Decisions may include such matters as (a) what diagnosis or diagnoses may be applicable, (b) what treatments are most likely to produce behavioral or emotional changes in desired directions, (c) what colleges should be considered, (d) what career options might be most satisfying, (e) whether an individual qualifies for a gifted educational program, (f) the extent to which an individual is at risk for given outcomes, (g) the extent to which an individual poses a risk of harm to others or to himself or herself, (h) the extent to which an individual has experienced deterioration in his or her ability to manage important aspects of living, and (i) whether an individual is suitable for particular types of roles or occupations such as those that involve high risk or extreme stress or where human error could have catastrophic effects. The foregoing list is certainly not exhaustive.

The term *assessment* as used in clinical and counseling settings is a broader term than *testing* because it refers to the more encompassing integration of information collected from numerous sources. Tests comprise sources of information that often contribute to assessment efforts. Discussion within this chapter focuses on procedures used in clinical and counseling assessment, all of which provide samples of behavior and, thus, qualify as tests. The narrative begins with a consideration of how clinical assessment may be framed and then addresses briefly ethics and other guidelines pertinent to assessment practices. Next, specific assessment techniques used in clinical and counseling contexts are reviewed, followed by a discussion of concerns related to interpretation and integration of assessment results. The chapter concludes with a section devoted to the importance of providing assessment feedback.

TRADITIONAL AND THERAPEUTIC ASSESSMENT

A diverse collection of procedures may be viewed as falling within the purview of clinical and counseling assessment (Naugle, 2009). The disparate array of procedures makes it somewhat difficult to appreciate commonalities among them, particularly for individuals who are relatively new to the field of assessment. Although clinical and counseling assessment procedures take many forms, nearly all are applied in a manner that facilitates an intense focus on concerns of a single individual or small unit of individuals, such as a couple or family (Anastasi & Urbina, 1997). The clinician who works one-on-one with a client during a formal assessment effectively serves as data collector, information processor, and clinical judge (Graham, 2006). Procedures that may be administered to groups of people often serve as screening measures that identify respondents who may be at

DOI: 10.1037/14048-001
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

risk and, therefore, need closer clinical attention (i.e., further testing conducted individually).

The immediate goals of clinical and counseling assessment frequently address mental illness and mental health concerns. Testing can help practitioners to better address an individual's mental illness or mental health needs by identifying those needs, improving treatment effectiveness, and tracking the process or progress of interventions (Carlson & Geisinger, 2009; Kubiszyn et al., 2000). Tests that assist clinicians' diagnostic efforts also may be important in predicting therapeutic outcome (i.e., prognosis) and establishing expectations for improvement. On a practical level, testing can be used to satisfy insurance or managed care requirements for evidence that supports diagnostic determinations or progress monitoring.

Within this basic framework, practitioners view the assessment process and their role within it differently. Indeed, some clinicians regard their role as similar to that of a technician or skilled tradesperson. From this traditional vantage point, skillful assessment begins to develop during graduate training, as trainees become familiar with the tools of the trade—tests, primarily. They learn about a variety of tests and how to use them. As trainees become practitioners, they accumulate experience with specific tests and find certain tests more helpful to their work with clients than other tests. It is not surprising that clinicians rely on tests that have proven most useful to them in their clinical work (Masling, 1992), despite test selection guidelines and standards that emphasize the importance of matching tests to the needs of the specific client or client's agent (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Eyde, Robertson, & Krug, 2010). As Cates (1999) observed, "the temptation to remain with the familiar [test battery] is an easy one to rationalize, but may serve the client poorly" (p. 637). It is important to note that the clinical milieu is fraught with immediate practical demands to provide client-specific information that is accurate, is useful, and addresses matters such as current conflicts, coping strategies, strengths and weaknesses, degree of distress, risk for self-harm, and so forth. The dearth of well-developed tests to assess certain clinical features does not alleviate or delay the need for this information in clinical practice. Thus, practitioners may find it necessary to do the best they can with the tools at hand.

Therapeutic assessment represents an alternative to traditional conceptualizations of the assessment process (Finn & Martin, 1997: Finn & Tonsager, 1997; Kubiszyn et al., 2000). In this contemporary framework, test givers and test takers collaborate throughout the assessment process and work as partners in the discovery process. Test takers have a vested interest in the initiation and implementation of assessment as well as in evaluating and interpreting results of the procedures used. Advocates of therapeutic assessment value and seek input from test takers throughout the assessment process and regard their perspectives as valid and informed. Rather than dismissing client input as fraught with self-serving motives and inaccuracies, practitioners who embrace the therapeutic assessment model engage clients as equal partners. This stance, together with the participatory role of the test giver, led Finn and Tonsager (1997) to characterize the process as an empathic collaboration in which tests offer opportunities for dialogue as well as interpersonal and subjective exchanges. A more thorough discussion of therapeutic assessment and its application is given in Chapter 26, this volume.

TEST USAGE

A survey of clinical psychology and neuropsychology practitioners (Camara et al., 2000) indicated that clinical psychologists most frequently used tests for personality or diagnostic assessment. The findings were consistent with those from an earlier study (O'Roark & Exner, 1989, as cited by Camara et al., 2000), in which 53% of psychologists also reported that they used testing to help determine the most effective therapeutic approach. Testing constitutes an integral component of many practitioners' assessment efforts as practitioners report using formal measures with regularity. Ball, Archer, and Imhof (1994) reported results from a national survey of a sample of 151 clinical psychologists who indicated they provided psychological testing services. The seven most used tests

reported by respondents were used by more than half of the practitioners who responded to the survey. In order, these tests included the Wechsler IQ scales, Rorschach, Thematic Apperception Test (TAT), Minnesota Multiphasic Personality Inventory (MMPI), Wide-Range Achievement Test, Bender Visual Motor Gestalt Test, and Sentence Completion. Camara et al.'s (2000) sample comprising 179 clinical psychologists reported remarkably similar frequencies of use, with the Wechsler IQ scales, MMPI, Rorschach, Bender Visual Motor Gestalt Test, TAT, and Wide-Range Achievement Test heading up the list. The preceding reports notwithstanding, considerable evidence suggests that test usage is in decline (Ben-Porath, 1997; Camara et al., 2000; Garb, 2003; Eisman et al., 2000; Meyer et al., 2001), whereas other researchers have noted a corresponding decline in graduate instruction and training in testing and assessment (Aiken, West, Sechrest, & Reno, 1990; Fong, 1995; Hayes, Nelson, & Jarrett, 1987).

The now ubiquitous presence of managed care in all aspects of health care, including mental health care, clearly influences practitioners' use of tests (Carlson & Geisinger, 2009; Yates & Taub, 2003). As is true for health care providers generally, mental health care providers can expect reimbursement for services they provide only if those services can be shown to be cost effective and essential for effective treatment. In a managed care environment, practitioners no longer have the luxury of making unilateral decisions about patient care, including test administration. Clinical assessments that pinpoint a diagnosis and provide direction for effective treatment are reimbursable, within limits, and typically are considered by third-party payers as therapeutic interventions (Griffith, 1997; Kubiszyn et al., 2000; Yates & Taub, 2003). Moreover, a number of studies have demonstrated that clinical tests have therapeutic value in and of themselves (Ben-Porath, 1997; Finn & Tonsager, 1997) and encourage their use as interventions.

STANDARDS, ETHICS, AND RESPONSIBLE TEST USE

Counseling and clinical psychologists who conduct assessments must maintain high standards and abide by recommendations for best practice. In short, their assessment practices must be beyond reproach. Considering the important and varied uses to which assessment results may be applied, it is not surprising that an array of rules, guidelines, and recommendations govern testing and assessment practices. For many years, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) have served several professions well as far as delineating the standards for test users as well as for test developers, and clinical and counseling psychologists must adhere to ethical principles and codes of conduct that influence testing practices.

The APA's *Ethical Principles of Psychologists and Code of Conduct* (*APA Ethical Principles*; APA, 2010) addresses assessment specifically in Standard 9, although passages relevant to assessment occur in several other standards, too. The 11 subsections of Standard 9 address issues such as use of tests, test construction, release of test data, informed consent, test security, test interpretation, use of automated services for scoring and interpretation, and communication of assessment results. In essence, the standards demand rigorous attention to the relationship between the clinician (as test giver) and the client (as test taker) from inception to completion of the assessment process. Ultimately, practitioners must select and use tests that are psychometrically sound, appropriate for use with the identified client, and responsive to the referral question(s). Furthermore, clinicians retain responsibility for all aspects of the assessment including scoring, interpretation and explanation of results, and test security, regardless of whether they choose to use other agents or services to carry out some of these tasks.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and Standard 9 of the *APA Ethical Principles* (APA, 2010) provide sound guidance for counseling and clinical psychologists who provide assessment-related services. A number of other organizations concerned with good testing practices have official policy statements that offer additional assistance to practitioners seeking further explication of testing-related guiding principles or whose services may extend to areas beyond traditional parameters. The policy statements most likely to interest counseling and clinical

psychologists include the *ACA Code of Ethics* (American Counseling Association, 2005), *Specialty Guidelines for Forensic Psychology* (Committee on the Revision of Specialty Guidelines for Forensic Psychology, 2011), *Principles for Professional Ethics* (National Association of School Psychologists, 2010), and the *International Guidelines for Test Use* (International Test Commission, 2001). In addition to the foregoing, many books about ethics in the professional practice of psychology include substantial coverage of ethical considerations in assessment (e.g., Cottone & Tarvydas, 2007; Ford, 2006). A particularly accessible volume by Eyde et al. (2010) provides expert analysis of case studies concerning test use in various settings, including mental health settings, and illustrating real-life testing challenges and conundrums.

ASSESSMENT METHODS

As in all assessment endeavors, tasks associated with assessment in clinical and counseling psychology involve information gathering. Clinical and counseling assessments typically comprise evaluations of individuals with the goal of assisting an individual client in some manner. To determine the best way to help an individual, clinicians rely on comprehensive assessments that evaluate several aspects of an individual's functioning. Thus, most such assessments involve collecting information using a variety of assessment techniques (e.g., interviews, behavioral observations). Moreover, the use of multiple procedures (e.g., tests) facilitates the overarching goal of clinical and counseling assessment and also resonates with the important principle of good testing practice. Specifically, Standard 11.20 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) states that, in clinical and counseling settings, "a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered" (p. 117). It follows that inferences drawn from a single measure must be validated against evidence derived from other sources, including other tests and procedures used in the assessment.

Counseling and clinical assessment methods vary widely in their forms. The means of identifying what information is needed and gathering relevant evidence may include direct communications with examinees, observations of examinees' behavior, input from other interested parties (e.g., family members, peers, coworkers, teachers), reviews of records (e.g., psychiatric, educational, legal), and use of formal measures (i.e., tests). Interviews, behavioral observations, and formal testing procedures represent the primary ways of obtaining clinically relevant information.

Interviewing

Intake or clinical interviews often represent a first point of contact between a client and a clinician in which information that contributes to clinical assessment surfaces. Many important concerns must be handled effectively within what is probably no more than a 50-minute session. Beyond practical (e.g., scheduling, billing, emergency contact information) and ethical (e.g., informed consent, confidentiality and its limits) matters, the practitioner must accurately grasp and convey his or her understanding of the issues to the client. If this understanding captures the client's concerns, then it likely helps the client to believe that his or her problems can be understood and treated by the clinician. If the practitioner's understanding of the client's issues is not accurate, then the client has the opportunity to provide additional information that represents his or her concerns more accurately. At the same time and somewhat in the background, the clinician exudes competence and concern in a manner that inspires hope and commitment, while, in the foreground, he or she establishes a fairly rapid yet accurate appraisal of the client's issues and concerns. Effective treatment depends on the establishment of rapport sufficient to suggest that a productive working relationship is possible along with an appraisal that accurately reflects the severity of the concerns expressed and disruptions in the client's ability to function on a day-to-day basis as well as attendant risks. For a more complete discussion, readers can consult Chapter 7, this volume, concerning clinical interviewing.

Many intake procedures involve clinical interviewing that is somewhat formalized by the use of a structured format or questionnaire. The quality of

intake forms varies widely, partly as a function of how they were developed. For example, clinicians may complete an intake form developed or adopted by the facility in which he or she works. Such forms generally include questions about the client's current concerns (e.g., "presenting problem" or "chief complaint") as well as historical information that may bear on the client's status (e.g., history of previous treatment, family history, developmental history). Depending on the quality of the intake form, practitioners may find it necessary to supplement the information collected routinely through completion of the form. In the appendices of her book, *The Beginning Psychotherapist's Companion,* Willer (2009) offers several lists of intake questions that may be used to probe specific areas of concern that may surface during the collection of intake information (e.g., depression and suicide, mania, substance use). Advisable in all clinical settings and essential in clinical settings that provide acute and crisis services, intake procedures must address the extent to which the client poses a danger to others or to himself or herself.

Intake interviews may be considered *semistructured* if they address specific content uniformly from one client to the next but are not tightly "scripted" as are *structured* interviews. According to Garb's (2005) review, semistructured interviews are more reliable than unstructured clinical interviews, most likely because of the similarity of content (if not actual test items) across interviewers. An example of a semistructured technique is the mental status examination (MSE), which refers to a standardized method of conducting a fairly comprehensive interview. The areas of mental status comprising an MSE are summarized in Table 1.1. Many MSE elements may be evaluated through unobtrusive observations made during the meeting or through verbal exchanges that occur naturally in ordinary conversation.

The semistructured nature of the MSE ensures coverage of certain vital elements of mental status but is flexible enough to allow clinicians to ask follow-up questions if he or she believes it is necessary or helpful to do so. The MSE is used by a wide variety of mental health providers (counseling and clinical psychologists as well as social workers, psychiatrists, and others) and typically is completed at intake or during the course of treatment to assess progress. There are several versions of the MSE, including standardized and nonstandardized forms (Willer, 2009). An example of a structured diagnostic interview is the Structured Clinical Interview for the *DSM–IV–TR* (SCID; First, Spitzer, Gibbon, & Williams, 2002), where *DSM–IV–TR* refers to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; American Psychiatric Association, 2000). Completion of the SCID allows practitioners to arrive at an appropriate psychiatric diagnosis.

Regardless of whether an initial clinical contact calls for formal assessment, a crucial area to evaluate during one's initial interactions with clients is the presence of symptoms that indicate risk of harm to self or others. "Assessing risk of suicide is one of the most important yet terrifying tasks that a beginning clinician can do" (Willer, 2009, p. 245) and constitutes the ultimate high-stakes assessment. It is also frequently encountered in clinical practice (Stolberg & Bongar, 2002). Multiple factors contribute to overall risk status either by elevating or diminishing risk. Bauman (2008) describes four areas to examine when evaluating risk of suicide: (a) short-term risk factors, including stressors arising from environmental sources and mental health conditions; (b) long-term precipitating risk factors, including genetic traits or predispositions and personality traits; (c) precipitating events, such as legal matters, significant personal or financial losses, unwanted pregnancy, and so forth; and (d) protective factors or buffers, such as hope, social support, and access to mental health services. An individual's overall risk of suicide represents a combination of risks emanating from the first three elements, which elevate overall risk, adjusted by the buffering effect of the last element, which reduces overall risk.

In practice, assessment of suicide risk relies heavily on clinical interviewing (Stolberg & Bongar, 2002). Specific tests designed to assess suicide risk, such as the Beck Hopelessness Scale (Beck, 1988) and the Suicide Intent Scale (Beck, Schuyler, & Herman, 1974), appear to be used infrequently by practitioners (Jobes, Eyman, & Yufit, 1990; Stolberg & Bongar, 2002). Assessment of risk must consider

TABLE 1.1

Major Areas Assessed During a Mental Status Examination

Area	Content
Appearance	The examiner observes and notes the person's age, race, gender, and overall appearance.
Movement	The examiner observes and notes the person's gait (manner of walking), posture, psychomotor excess or retardation, coordination, agitation, eye contact, facial expressions, and similar behaviors.
Attitude	The examiner notes client's overall demeanor, especially concerning cooperativeness, evasiveness, hostility, and state of consciousness (e.g., lethargic, alert).
Affect	The examiner observes and describes affect (outwardly observable emotional reactions), as well as appropriateness and range of affect.
Mood	The examiner observes and describes mood (underlying emotional climate or overall tone of the client's responses).
Speech	The examiner evaluates the volume and rate of speech production, including length of answers to questions, the appropriateness and clarity of the answers, spontaneity, evidence of pressured speech, and similar characteristics.
Thought content	The examiner assesses what the client says, listening for indications of evidence of misperceptions, hallucinations, delusions, obsessions, phobias, rituals, symptoms of dissociation (feelings of unreality, depersonalization), or thoughts of suicide.
Thought process	The examiner assesses thought processes (logical connections between thoughts and how thoughts connect to the main thread or gist of conversation), noting especially irrelevant detail, verbal perseveration, circumstantial thinking, flight of ideas, interrupted thinking, and loose or illogical connections between thoughts that may indicate a thought disorder.
Cognition	The evaluation assesses the person's orientation (ability to locate himself or herself) with regard to person, place, and time; long- and short-term memory; ability to perform simple arithmetic (e.g., serial sevens); general intellectual level or fund of knowledge (e.g., identifying the last several U.S. presidents, or similar questions); ability to think abstractly (explaining a proverb); ability to name specific objects and read or write complete sentences; ability to understand and perform a task with multiple steps (e.g., showing the examiner how to brush one's teeth, throw a ball, or follow simple directions); ability to draw a simple map or copy a design or geometrical figure; ability to distinguish between right and left.
Judgment	The examiner asks the person what he or she would do about a commonsense problem, such as running out of shampoo.
Insight	The examiner evaluates degree of insight (ability to recognize a problem and understand its nature and severity) demonstrated by the client.
Intellectual	The examiner assesses fund of knowledge, calculation skills (e.g., through simple math problems), and abstract thinking (e.g., through proverbs or verbal similarities).

several features of risk beyond its mere presence including immediacy, lethality, and intent. *Immediacy* represents a temporal consideration with higher levels of immediacy associated with imminent risk—a state of acute concern for the individual's life. Assessment of imminent risk involves consideration of several empirically derived risk factors including (a) history of prior attempts (with recent attempts given greater weight than attempts that occurred longer ago); (b) family history of suicide or attempt; and (c) presence of mental or behavior disorders such as substance abuse, depression, and conduct disorder. Imminent risk is accelerated by an inability to curb impulses and a need to "blow off steam," which constitute poor prognostic signs. *Lethality* refers to the possibility of death occurring as a result of a particular act. In assessing risk of suicide, the act in question is one that is planned or contemplated by the client. Use of firearms connotes higher lethality than overdosing on nonprescription drugs (e.g., aspirin). Lethality differs from *intent,* which refers to what the person seeks to accomplish with a particular act of self-harm. Serious suicidal intent is not necessarily associated with acts of high lethality.

Behavioral Observations

One of the earliest means by which assessment information begins to accumulate is the test taker's

behaviors. Surprisingly little information about behavioral observations appears in the empirical or practice-based literature, despite its traditional inclusion as a section in assessment reports (Leichtman, 2002; Tallent, 1988). Although difficult to standardize and quantify, many psychologists consider the observations and interpretations of an examinee's behavior during testing vital to understanding the client (Oakland, Glutting, & Watkins, 2005). Only a few standardized assessments of test behavior have been developed, sometimes associated with a specific test. For example, Glutting and Oakland (1993) developed the Guide to the Assessment of Test Session Behavior and normed it on the standardization samples of the Wechsler Intelligence Scale for Children (3rd ed.; Wechsler, 1993) and the Wechsler Individual Achievement Test (Psychological Corporation, 1992). To date, standardized measures of test session behavior have not been widely adopted.

Counseling and clinical psychologists typically have sufficient and specialized training to allow them to observe and record an examinee's verbal and nonverbal behaviors. Notations usually are made for several behavioral dimensions including physical appearance, attitude toward testing, content of speech, quality and amount of motor activity, eye contact, spontaneity, voice quality, effort (generally and in the face of challenge), fatigue, cooperation, attention to tasks, willingness to offer guesses (if applicable), and attitude toward success and failure (if applicable). Leichtman (2002) cautioned against either (a) including observations of everything a test taker thinks, feels, says, and does; or (b) reducing behavioral descriptions to such an extent that the resulting narrative fails to provide any real sense of what the test taker is like.

Behavior during clinical and counseling testing is unavoidably influenced by interactions between the test taker and the test giver. As Masling (1992) observed, "the psychologist is simultaneously a participant in the assessment process and an observer of it" (p. 54). A common expectation and responsibility of psychologists who administer such tests is to establish rapport with the test taker before implementing test procedures. Rapport is vital to ensure a test taker's cooperation and best effort, attitudes that contribute to test results that provide an accurate portrayal of the test taker's characteristics. However, rapport differs from one dyad to another, as stylistic and personality factors vary across both examiners and examinees and affect the quality of their interactions. Although adherence to standardized administration procedures during testing is vital to preserve the integrity of assessment process and test score interpretability (e.g., AERA, APA, & NCME, 1999; Geisinger & Carlson, 2009), practitioners are not automatons who simply set specific tasks before examinees while reciting specific instructions. Actions taken by examiners during individual test administration must be responsive to test-taker behaviors and the examiner's interpretation of those behaviors. Some of these actions are scripted in the test administration procedures, whereas others are subtle, nonverbal—possibly unconscious—ones that serve to allay anxiety or encourage elaboration of a response. Other actions follow logically from an examinee's behavior, such as when the examiner offers a short break after noting the examinee's failed attempt to stifle several yawns. In this vein, Leichtman (2002) suggested that test administration procedures and instructions are "like a play. Examiners are bound by the script, but there is wide latitude for how they and their clients interpret their roles" (p. 209). The traditions of testing encourage the notion that an examiner, "like the physical scientist or engineer, is 'measuring an object' with a technical tool. But the 'object' before him [sic] is a person, and the testing involves a complex psychological relationship" (Cronbach, 1960, p. 602).

Formal Testing

Tests are used by counseling and clinical psychologists at various points in therapeutic contexts. Some tests may be administered during an intake session, before the establishment of a therapeutic relationship, to check for a broad range of possible issues that may need clinical attention. These screening measures represent a "first pass" over the variety of issues that may concern a person who seeks mental health assistance. They are meant to provide a gross indication of level of symptom severity in select areas and, often, to indicate where to focus subsequent assessment efforts (Kessler et al., 2003).

Screening measures typically are quite brief and are seldom, if ever, validated for use as diagnostic instruments. Rather, these measures provide a glimpse into the nature and intensity of a client's concerns. As such, they may reveal problems that need immediate attention as well as areas needing further assessment. An example of a screening measure designed for use in college counseling centers is the Inventory of Common Problems (ICP; Hoffman & Weiss, 1986), a 24-item inventory of specific problems college students may encounter. Respondents use a 5-point Likert-type scale to indicate the extent to which they have been bothered or worried by the stated problem over the past few weeks. Areas assessed include depression, anxiety, academic problems, interpersonal problems, physical health problems, and substance use problems. High scores suggest topics that may be explored further in counseling.

The Symptom Check List-90-R (SCL-90-R; Derogatis, 1994) is a clinical screening inventory with broader applicability than the ICP. The inventory consists of 90 items, each of which presents a symptom of some sort to which respondents indicate the extent to which they were distressed by that symptom over the past week, using a 5-point scale. The SCL-90-R yields scores on nine scales (Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism) and total scores on three scales (Global Severity Index, Positive Symptom Total, and Positive Symptom Distress Index). Norms are differentiated by age (adolescent and adult) for nonpatients and by psychiatric patient status (nonpatient, inpatient, and outpatient) for adults, with each norm keyed by gender. Some brief clinical measures may be used to screen for problems in a single area of potential concern. For example, the Beck Depression Inventory—II (Beck, Steer, & Brown, 1996) and the State–Trait Anxiety Inventory (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) screen for elevated levels of symptom severity in depression and anxiety, respectively. Overall, these and other screening measures are most useful for detecting cases in need of further examination.

The assessment procedures described thus far are used routinely at or near the outset of a therapeutic relationship to help specify or clarify the clinical situation that prompted the client to seek treatment. More extensive, formal testing may prove beneficial at an early stage of intervention or anytime during therapy to specify, clarify, or differentiate diagnoses; to monitor treatment progress; or to predict psychotherapy or mental health outcomes (Kubiszyn et al., 2000; see also Chapter 13, this volume, concerning psychological assessment in treatment). Counseling and clinical testing can be used to illuminate a variety of dimensions that may help clinicians to deliver effective treatment for a particular client, including measures of cognitive ability, values, interests, academic achievement, psychopathology, personality, and attitudes. The sheer number of tests available in each of these areas makes it impractical to review (or even mention) every test that may have clinical salience, particularly in light of the coverage afforded these measures in other chapters of this handbook. Thus, in the section that follows, tests are described according to several different ways of grouping them, with implications for clinical and counseling tests highlighted.

DIMENSIONS OF CLINICAL AND COUNSELING TESTING

Various characteristics of tests may be used to distinguish among them. Such distinctions go beyond merely grouping or categorizing tests. For example, tests differ in administration format, nature of the respondent's tasks, and whether the stakes associated with the use of test scores are high or low. These dimensions influence the testing process in counseling and clinical contexts, by affecting expectations and behaviors of test givers and test takers as well as how the tests may be used and the confidence testing professionals may have in the results.

Test administration format is one way to distinguish among tests. Some tests require one-to-one or individual administration, whereas other tests are designed for group administration. Generally speaking, it is possible to administer group tests using an individual format, although the examiner's role in these situations is often reduced as he or she serves primarily as a monitor of the session. As suggested near the beginning of this chapter, clinical measures

focus intensely on individual concerns. It follows that many—although by no means all—clinical measures were developed for individual administration. Individually administered tests are highly dependent on the clinical skills of the examiner. As Meyer et al. (2001) observed, "a psychological test is a dumb tool, and the worth of the tool cannot be separated from the sophistication of the clinician who draws inferences from it and then communicates with patients and other professionals" (p. 153). Among other things, the responsibility to establish and maintain rapport rests with the clinician, and there is no magic formula by which to achieve it and no established criteria by which to establish that a reasonable level of rapport has been achieved. That determination depends on clinical judgment.

At the outset of a testing session, examiners need to ensure that a sufficient level of comfort and communication exists with the test taker to foster his or her best and sustained effort. Examiners need to exude a businesslike manner yet remain responsive to queries from the test taker and aware of fluctuation in the test taker's energy, focus, and attitude. They need to help test takers understand that testing is important but must avoid overstating this point, lest the test taker become overly anxious about performing well on the test tasks. Test takers differ in terms of their readiness to engage in the assessment process and to give it their best effort: Some are eager to begin, some are anxious, some are irritated, some are suspicious or confused, and so forth. The clinician must keep a finger on the pulse of the testing session and take action as needed to restore rapport and keep motivation high and performance optimal.

Standardized individual administration of tests is vital for the vast majority of tests to assure that testing conditions are the same for all test takers; therefore, results from different test takers may be meaningfully compared (Geisinger & Carlson, 2009). However, given the interpersonal context within which clinical and counseling measures are administered, this procedural sameness is difficult to ensure for all aspects of testing. For example, most projective (performance-based) measures are untimed. How long examiners wait before moving on to the next stimulus is a matter of judgment and, likely, varies a great deal from one examiner to the next. Some standardized measures include "scripts" for the examiner, in an effort to make administration more uniform across examiners. Despite appearances, there is room for interpretation in the scripts nevertheless (Leichtman, 2002). How scrupulously examiners follow standardized procedures for administration is an open question (Geisinger & Carlson, 2009; Masling, 1992), as studies of even highly scripted individually administered tests reveal many departures (e.g., Moon, Blakey, Gorsuch, & Fantuzzo, 1991; Slate, Jones, & Murray, 1991; Thompson & Bulow, 1994).

On the other hand, group-administered tests are not monitored as closely as individually administered tests and do not depend on rapport to ensure optimal performance. Directions for group-administered tests must be clear to all test takers before the beginning of the test (or inventory or questionnaire) because missteps by examinees cannot be corrected easily. The same instructions and practice procedures are used for everyone. An individual who perhaps would benefit from one more practice item will not get it, and there will be no follow-up opportunities to test limits.

The nature of the tasks that constitute individual tests is another way to distinguish tests. In Chapter 10 of this volume, which addresses performance-based measures (often referred to as *projective techniques*), Irving B. Weiner describes a major distinction between test types—that is, between performance-based measures and self-report measures. The former test type requires test takers to act upon stimuli presented to them (e.g., Rorschach inkblots, TAT cards), to create or construct responses, or to formulate responses to specific questions (e.g., Wechsler scales of intelligence) presented to them, whereas self-report measures ask respondents to answer questions about themselves by selecting responses from a preset array of options. As suggested by Weiner, neither test type is inherently superior, as the test types seek and provide different kinds of information. A test's clinical value is unrelated to the nature of the tasks that constitute it.

Performance-based measures typically use scoring systems or rubrics that ultimately depend on

some degree of subjectivity in scoring. The tasks that constitute performance-based measures are open-ended and offer wide latitude to test takers as far as how they choose to respond. Some tests or tasks require constructed responses (e.g., TAT, figure drawings), whereas others require retrieval or application of specific information (e.g., Vocabulary and Arithmetic subtests on the Wechsler tests).

Self-report measures require examinees to select or endorse a response presented in a predefined set of possibilities. In part because responses are selected rather than constructed by the examinee, systematic distortion of responses is a concern in many self-report inventories (Graham, 2006). Detecting such response sets is important because, when they occur, they may undermine the validity of the test scores. Validity scales were big news when they were first introduced in the original MMPI (Hathaway & McKinley, 1943); now they are commonplace in many personality and other types of inventories. Scoring of self-report measures is considered to be objective and typically involves the use of either computer software or scoring templates. Other than human errors (e.g., misaligning a scoring template), objective scoring produces test scores that do not require clinical judgment. Detailed discussion of self-report measures is provided later in Chapter 11, this volume.

The level of impact that the use of tests scores may have varies and forms another way to distinguish groups of tests. *High-stakes testing* refers to the situation where test scores are used to make important decisions about an individual. The impact level of such decisions is substantial, sometimes rising to the level of life altering. Tests whose results are used to render such decisions must be psychometrically sound. Evidence supporting the reliability and validity of test scores must surpass the level typically seen in measures used for lesser purposes, such as research or screening. Custody evaluations used to determine parental fitness (for further information, see Chapter 34, this volume) and forensic evaluations used to establish competency to stand trial (for further information, see Chapters 6 and 16, this volume) are but two examples of high-stakes testing situations.

In clinical decision making, the specific test used does not automatically determine the stakes. Rather, the use to which the test scores are put dictates whether the testing should be considered high stakes. For example, practitioners may use the results of an assessment simply to confirm a diagnosis and formulate interventions. This use of tests is a rather routine practice aimed at improving the mental health of a particular client. In this situation, the stakes likely are low, because the individual is already engaged in treatment and the differential diagnosis that is sought will enhance the clinician's understanding and treatment of his or her psychological difficulties. If the same test results were used as the basis for denying disability benefits, then the testing context would be regarded as high stakes.

Low-stakes measures often include those related to documenting values and interests. The human interest value of these measures notwithstanding, low-stakes situations simply do not have the same level of impact as high-stakes decisions. Test takers frequently are curious to review the assessment results, but many are not surprised by them. However, low-stakes measures may contribute to important decisions that an individual may make concerning career or relationship pursuits or other quality-of-life choices.

INTERPRETING AND INTEGRATING ASSESSMENT RESULTS

Interpreting and integrating test results requires a tenacious, disciplined, and thorough approach. It follows the collection of data from various sources, none of which should be ignored or dismissed. Like test administration, test interpretation represents

> an interpersonal activity [that] may be considered part of the influence process of counseling. The counselor communicates his or her own understanding of the client's test data to the client, anticipating that the client will adopt and apply some part of that understanding as self-understanding. (Claiborn & Hanson, 1999, p. 151)

An important objective in interpreting assessment results is to account for as much test data as possible. Formulating many tenable hypotheses at

the outset of test interpretation facilitates this goal. With regard to enhancing clinical judgment, Garb (1989) encouraged clinicians to become more willing to consider alternative hypotheses and to revise their initial views of a client's behavior. Although Garb's point referred broadly to clinical judgment and not specifically to clinical assessment, it applies equally well to test interpretation. For example, an overarching ennui reported by an adult client at intake could stem from numerous causes, including psychological and physical ones. Subsequent results from a comprehensive assessment consisting of a multitude of tests and sources of data may suggest (a) depression or a related derivative, (b) bereavement, (c) malingering, (d) anemia, (e) reaction to situational (e.g., job related) stress, (f) passive–aggressive coping strategy, (g) insomnia, (h) a side effect of a new medication, (i) a combination of two or more of the foregoing, or (j) something else entirely. An intake interview and routine screening measures may rule out several of the possible explanations. Interpretations stemming from more comprehensive measures may be compared against the remaining competing hypotheses to ascertain which hypothesis best accounts for the evidence. In the end, the best explanation is the one that explains most (or all) of the evidence accumulated and considered in the assessment process.

An important first step in evaluating test data often takes place while assessment procedures are under way, in the presence of the test taker or before he or she leaves the premises where testing occurred. This step involves reviewing the examinee's responses to any "critical items" that are included on any of the measures. These items are so called because their content has been judged to be indicative of serious maladjustment, signifying grave concerns such as the propensity for self-harm. Although empirical scrutiny has not tended to offer much support for the utility of critical items for this purpose (Koss, 1980; Koss, Butcher, & Hoffman, 1976), many practitioners consider the items worthy of follow-up efforts, perhaps because failing to act on such a blatant appeal for assistance would be unconscionable and the possible outcome irreversible. Moreover, base-rate problems cloud the issue, as low-base-rate events such as suicide are notoriously difficult to predict (Sandoval, 1997), especially when one tries to predict such an event on the basis of responses to a small handful of items. Also at issue is the absence of an adequate criterion against which to judge test validity (Hunsley & Meyer, 2003). A client who does not commit suicide after his or her responses to critical items suggested a high risk of suicide was present was not necessarily misjudged. Individuals at high risk for a given outcome do not unerringly suffer that outcome; such is the nature of risk.

Base-rate and criterion problems persist in the area of suicide risk assessment and are unlikely to be resolved. Measures developed to assess suicide risk are intended to be used to avert acts of self-harm and cannot be easily validated in the usual manner because lives are at stake. Critical items denote risk; they do not predict behavior. Recommended practice is to avoid treating critical items as a scale or brief assessment of functioning, but rather consider the items as offering possible clues to content themes that may be important to the client (Butcher, 1989).

After considering a client's responses to critical items, integration of findings obtained from the various methods used in an assessment moves to a review of evidence collected during the assessment, including test and nontest data, from each individual source. Scoring and interpreting or evaluating individual procedures that were implemented constitutes an important first step because it is at this stage that clinicians begin to weigh the credibility of the evidence. Specifically, it is essential to note for each procedure whether the test taker's approach to that procedure allows further consideration of the results. Tests that include validity scales can make this task more objective and fairly straightforward. However, many assessment procedures do not have built-in components to help examiners evaluate whether responses should be considered valid indicators of the test taker's functioning. In these cases, examiners must render a judgment, often based on the test taker's demeanor, attitude toward the procedures, and behaviors demonstrated during the assessment. Obviously and unfortunately, this judgment process is not standardized and is quite open to subjective interpretations. Even so, it is probably safe to conclude that most practitioners would at

least question the validity of assessment results from a client who arrived to the session 20 minutes late, looked at his watch no fewer than 25 times, neglected to respond to half of the items on two test forms, and sighed audibly throughout the assessment while mumbling about how "ridiculous this is." In any case,

> psychologists must consider whether there is a discernible reason for test takers to be less than forthright in their responses, and whether that reason might constitute a motive for faking. If so, the test giver must . . . interpret test findings with these possibilities in mind. (Carlson & Geisinger, 2009, p. 83)

In the early stages of interpretation, possible explanations for the results should be treated as tentative, because various hypotheses may be offered to explain individual test outcomes. All reasonable explanations for the observed results should be considered while examining evidence from other sources. In the face of additional data, some hypotheses will be discarded and some will be retained. Evidence from other sources—test and nontest—that confirms or disconfirms active hypotheses is particularly important, as this type of evidence helps to bolster (i.e., rule in) or weaken (i.e., rule out) putative explanations, respectively. Typically, a small number of hypotheses survive this iterative process, and these viable explanations of the observed results form the prominent themes of a written report.

PROVIDING ASSESSMENT FEEDBACK

Providing test feedback to test takers is an ethical responsibility (e.g., APA, 2010) that appears to be taken lightly by some practitioners according to some published reports (Pope, 1992; Smith, Wiggins, & Gorske, 2007). As Smith et al. (2007) observed, there is surprisingly little written about assessment feedback and "little published research on the assessment feedback practices of psychologists" (p. 310). These researchers surveyed some 719 clinicians (neuropsychologists and members of the Society for Personality Assessment) about their psychological assessment feedback practices to find that some 71% reported that they frequently provided in-person feedback, either to clients or clients' family members. The researchers also queried respondents about the time they spent providing feedback, how useful they found the practice, and what kind of feedback they provided (e.g., written, oral). Although most practitioners reported that they do provide feedback, nearly 41% reported that they provided no direct feedback to clients or their families. Nearly one third of respondents reported that they mailed a report to clients, a practice that Harvey (1997) denounced, because recipients often lack the background and technical knowledge to understand and interpret the results. Even so, Smith et al. viewed the survey results positively overall and suggested that the status of psychological assessment feedback practices may not be as dire as suggested several years ago (Pope, 1992). Interested readers may refer to Chapter 3, this volume, for further guidance on communicating assessment results.

Test feedback may serve several important purposes, not the least of which is to help bring about behavioral changes (Butcher, 2010; Finn & Tonsager, 1997). In discussing the importance of providing test feedback, Pope (1992) suggested that the feedback process offers opportunities on several fronts that bear directly on the therapeutic process and that, in essence, extend the assessment to include the feedback component. Empirical evidence accumulated thereafter, which demonstrated treatment effects of assessment feedback (Kubiszyn et al., 2000). Specifically, several studies compared therapeutic gains made by clients in treatment who received feedback about their test results on the MMPI–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) to those of similar clients who did not receive such feedback (e.g., Finn & Tonsager, 1992, 1997; Fischer, 2000; Newman & Greenway, 1997). Clients who received assessment feedback demonstrated therapeutic improvements, as noted by their higher levels of hope and decrease in reported symptoms.

CONCLUDING THOUGHTS

Assessment methods used in counseling and clinical contexts focus tightly on an individual client's condition and seek to identify ways in which his or

her concerns may be addressed or resolved. Broadly speaking, the methods used include interview techniques, behavioral observations, and formal tests that place different demands on the examinee as well as the examiner. Information gathered from multiple sources then must be interpreted and integrated into a cohesive explanation of the test data and, by extension, the client's functioning and features. The end goal of assessment in counseling and clinical contexts is to produce an accurate portrayal of the client's functioning that is useful for planning and implementing interventions. Providing feedback to the client about assessment results is vital to promoting the client's interests and effecting treatment.

Cates (1999) observed that clinical assessment is best regarded as providing a "snapshot not a film" of an individual's functioning, that "describes a moment frozen in time, described from the viewpoint of the psychologist" (p. 637). When an observer says something like, "that's a good picture of her," the speaker means that the image represents the subject as she truly is. Good pictures depend on using good tools and good techniques. Clinical assessment, too, uses tools and techniques to reflect the characteristics of the client as he or she exists and functions every day.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, P. R. (1990). Graduate training in statistics, methodology and measurement in psychology: A survey of Ph.D. programs in North America. *American Psychologist, 45*, 721–734. doi:10.1037/0003-066X.45.6.721

American Counseling Association. (2005). *ACA code of ethics*. Washington, DC: Author. Retrieved from http://72.167.35.179/Laws_and_Codes/ACA_Code_of_Ethics.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Ball, J. D., Archer, R. P., & Imhof, E. A. (1994). Time requirements of psychological testing: A survey of practitioners. *Journal of Personality Assessment, 63*, 239–249. doi:10.1207/s15327752jpa6302_4

Bauman, S. (2008). *Essential topics for the helping professional*. Boston, MA: Pearson.

Beck, A. T. (1988). *Beck Hopelessness Scale*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Schuyler, D., & Herman, I. (1974). Development of suicidal intent scales. In A. T. Beck, H. L. P. Resnik, & D. J. Lettieri (Eds.), *The prediction of suicide* (pp. 45–56). Bowie, MD: Charles Press.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Ben-Porath, Y. S. (1997). Use of personality instruments in empirically guided treatment planning. *Psychological Assessment, 9*, 361–367. doi:10.1037/1040-3590.9.4.361

Butcher, J. N. (1989). *The Minnesota report: Adult Clinical System MMPI–2*. Minneapolis, MN: University of Minnesota Press.

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. doi:10.1146/annurev.clinpsy.121208.131420

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for administration and scoring: Minnesota Multiphasic Personality Inventory—2 (MMPI–2)*. Minneapolis, MN: University of Minnesota Press.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi:10.1037/0735-7028.31.2.141

Carlson, J. F., & Geisinger, K. F. (2009). Psychodiagnostic testing. In R. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 67–88). Washington, DC: American Psychological Association. doi:10.1037/11861-002

Cates, J. A. (1999). The art of assessment in psychology: Ethics, expertise, and validity. *Journal of Clinical Psychology, 55*, 631–641. doi:10.1002/(SICI)1097-4679(199905)55:5<631::AID-JCLP10>3.0.CO;2-1

Claiborn, C. D., & Hanson, W. E. (1999). Test interpretation: A social-influence perspective. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 151–166). Needham Heights, MA: Allyn & Bacon.

Committee on the Revision of the Specialty Guidelines for Forensic Psychology. (2011). *Specialty guidelines for forensic psychology* (6th draft). Retrieved from http://www.ap-ls.org/aboutpsychlaw/3182011sgfpdraft.pdf

Cottone, R. R., & Tarvydas, V. M. (2007). *Counseling ethics and decision-making* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper.

Derogatis, L. R. (1994). *Administration, scoring, and procedures manual for the SCL-90-R*. Minneapolis, MN: National Computer Systems.

Eisman, E. J., Dies, R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., . . . Moreland, K. L. (2000). Problems and limitations in the use of psychological assessment in contemporary health care delivery. *Professional Psychology: Research and Practice*, *31*, 131–140. doi:10.1037/0735-7028.31.2.131

Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior* (2nd ed.). Washington, DC: American Psychological Association.

Finn, S. E., & Martin, H. (1997). Therapeutic assessment with the MMPI–2 in managed health care. In J. N. Butcher (Ed.), *Objective personality assessment in managed health care: A practitioner's guide* (pp. 131–152). New York, NY: Oxford University Press.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI–2 test feedback to college students awaiting therapy. *Psychological Assessment, 4*, 278–287. doi:10.1037/1040-3590.4.3.278

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9*, 374–385. doi:10.1037/1040-3590.9.4.374

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (2002). *Structured Clinical Interview for* DSM–IV–TR *Axis I disorders, research version, patient edition (SCID-I/P)*. New York, NY: Biometrics Research, New York State Psychiatric Institute.

Fischer, C. T. (2000). Collaborative, individualized assessment. *Journal of Personality Assessment, 74*, 2–14. doi:10.1207/S15327752JPA740102

Fong, M. L. (1995). Assessment and *DSM–IV* diagnosis of personality disorders: A primer for counselors. *Journal of Counseling and Development, 73*, 635–639. doi:10.1002/j.1556-6676.1995.tb01808.x

Ford, G. G. (2006). *Ethical reasoning for mental health professionals*. Thousand Oaks, CA: Sage.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*, 387–396. doi:10.1037/0033-2909.105.3.387

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15*, 508–520. doi:10.1037/1040-3590.15.4.508

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1*, 67–89. doi:10.1146/annurev.clinpsy.1.102803.143810

Geisinger, K. F., & Carlson, J. F. (2009). Standards and standardization. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 99–111). New York, NY: Oxford University Press.

Glutting, J., & Oakland, T. (1993). *Guide to the assessment of test session behavior: Manual*. San Antonio, TX: Psychological Corporation.

Graham, J. R. (2006). *MMPI–2: Assessing personality and psychopathology* (4th ed.). New York, NY: Oxford University Press.

Griffith, L. (1997). Surviving no-frills mental health care: The future of psychological assessment. *Journal of Practical Psychiatry and Behavioral Health, 3*, 255–258.

Harvey, V. S. (1997). Improving readability of psychological reports. *Professional Psychology: Research and Practice, 28*, 271–274. doi:10.1037/0735-7028.28.3.271

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42*, 963–974. doi:10.1037/0003-066X.42.11.963

Hoffman, J. A., & Weiss, B. (1986). A new system for conceptualizing college students' problems: Types of crises and the Inventory of Common Problems. *Journal of American College Health, 34*, 259–266. doi:10.1080/07448481.1986.9938947

Hood, A. B., & Johnson, R. W. (2007). *Assessment in counseling: A guide to the use of psychological assessment procedures*. Alexandria, VA: American Counseling Association.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455. doi:10.1037/1040-3590.15.4.446

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*, 93–114. doi:10.1207/S15327574IJT0102_1

Jobes, D. A., Eyman, J. R., & Yufit, R. I. (1990, April). *Suicide risk assessment survey*. Paper presented at the annual meeting of the American Association of Suicidology, New Orleans, LA.

Kessler, R. C., Barker, P. R., Cople, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., . . . Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry*, *60*, 184–189. doi:10.1001/archpsyc.60.2.184

Koss, M. P. (1980). *Assessment of psychological emergencies with the MMPI*. Nutley, NJ: Roche.

Koss, M. P., Butcher, J. N., & Hoffman, N. (1976). The MMPI critical items: How well do they work? *Journal of Consulting and Clinical Psychology, 44*, 921–928. doi:10.1037/0022-006X.44.6.921

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., . . . Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice*, *31*, 119–130. doi:10.1037/0735-7028.31.2.119

Leichtman, M. (2002). Behavioral observations. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 303–318). New York, NY: Oxford University Press.

Masling, J. M. (1992). Assessment and the therapeutic narrative. *Journal of Training and Practice in Professional Psychology, 6*, 53–58.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Moon, G. W., Blakey, W. A., Gorsuch, R. L., & Fantuzzo, J. W. (1991). Frequent WAIS–R administration errors: An ignored source of inaccurate measurement. *Professional Psychology: Research and Practice, 22*, 256–258. doi:10.1037/0735-7028.22.3.256

National Association of School Psychologists. (2010). *Principles for professional ethics*. Retrieved from http://www.nasponline.org/standards/2010standards/1_%20Ethical%20Principles.pdf

Naugle, K. A. (2009). Counseling and testing: What counselors need to know about state laws on assessment and testing. *Measurement and Evaluation in Counseling and Development, 42*, 31–45. doi:10.1177/0748175609333561

Newman, M. L., & Greenway, P. (1997). Therapeutic effects of providing MMPI–2 test feedback to clients at a university counseling service: A collaborative approach. *Psychological Assessment, 9*, 122–131. doi:10.1037/1040-3590.9.2.122

Oakland, T., Glutting, J., & Watkins, M. W. (2005). Assessment of test behaviors with the WISC–IV. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC–IV clinical use and interpretations: Scientist-practitioner perspectives* (pp. 435–467). San Diego, CA: Elsevier Academic Press.

Pope, K. S. (1992). Responsibilities in providing psychological test feedback to clients. *Psychological Assessment, 4*, 268–271. doi:10.1037/1040-3590.4.3.268

Psychological Corporation. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Author.

Sandoval, J. (1997). Critical thinking in test interpretation. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 31–49). Washington, DC: American Psychological Association.

Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale—Revised: An empirical evaluation of practice administrations. *Professional Psychology:, Research and Practice, 22*, 375–379. doi:10.1037/0735-7028.22.5.375

Smith, S. R., Wiggins, C. M., & Gorske, T. T. (2007). A survey of psychological assessment feedback practices. *Assessment, 14*, 310–319. doi:10.1177/1073191107302842

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for State–Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Stolberg, R., & Bongar, B. (2002). Assessment of suicide risk. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 376–406). New York, NY: Oxford University Press.

Tallent, N. (1988). *Psychological report writing* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Thompson, A. P., & Bulow, C. A. (1994). Administration error in presenting the WAIS–R blocks: Approximating the impact of scrambled presentations. *Professional Psychology: Research and Practice, 25*, 89–91. doi:10.1037/0735-7028.25.1.89

Wechsler, D. (1993). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: Psychological Corporation.

Willer, J. (2009). *The beginning psychotherapist's companion*. Lanham, MD: Rowman & Littlefield.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost–benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15*, 478–495. doi:10.1037/1040-3590.15.4.478

CHAPTER 2

THE ASSESSMENT PROCESS

Sara Maltzman

This chapter reviews the historical purposes of psychological assessment, the components and process of psychological assessment, current issues, and emerging trends. In keeping with the emphases of this handbook, the discussion focuses on the use of assessments and the assessment process within clinical, counseling, and forensic psychology.

THE HISTORY OF PSYCHOLOGICAL ASSESSMENTS

McGuire (1990) traced the development of formal psychological testing to James McKean Cattell in the 1890s and early 20th century. McGuire noted that Cattell and the first few experimental psychologists who came to define themselves as clinical psychologists advocated for education, training, and the establishment of professional standards for the assessment of intellectual and personality functioning. Thus, the assessment and diagnosis of intellectual functioning and personality were the fundamental functions of clinical psychologists. Witmer, who made significant contributions to the development of clinical, developmental, and educational psychology, established the first psychological clinic in 1896 (Baker, 1988). The clinic assessed and treated children who presented with possible mental retardation, learning disabilities, or emotional concerns that prevented attainment of their academic potential. Witmer utilized a multidimensional, functional approach that included a comprehensive psychosocial history taking as well as behavioral observations in multiple environments (e.g., home, school) over time. A physician completed the physical examination, and often the behavioral observations were made by a social worker. These data were summarized into an integrative assessment of the child's deficiencies, along with treatment recommendations (Baker, 1988). Thus, a primary focus within clinical psychology at the beginning of the 20th century was the multimodal assessment, diagnosis, and treatment of children and youths.

The treatment recommendations made for these youths often included vocational direction (Baker, 2002). With the stock market crash and high unemployment of the 1930s, the vocational needs of adults began to predominate and the vocational assessment of youths transitioned to adult vocational counseling and later into the field of counseling psychology for adults (Baker, 2002; Super, 1955). The assessment of aptitudes as well as of abilities emerged out of the necessity to assist the unemployed. At the same time, Rogerian theory and its associated nondirective, client-centered therapeutic approach began to emerge. The Rogerian approach was applied to vocational counseling in recognition that such an orientation was theoretically compatible with counseling focused on the achievement of vocational aspirations (Super, 1955). These three foci—the assessment of aptitudes, the assessment of abilities, and a Rogerian conceptualization of the person and the therapeutic relationship—converged into a cohesive approach for addressing the psychosocial concerns of the unemployed. Over time, this approach was modified to address the needs of returning World War II (WWII) veterans and to

DOI: 10.1037/14048-002
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

assist them in maximizing their psychosocial strengths. Addressing the vocational, educational, and adjustment needs of returning WWII veterans led to the establishment of counseling psychology as a distinct position within the U.S. Veterans Administration (VA) system (Meara & Myers, 1999). To meet the needs of returning veterans, the VA encouraged the American Psychological Association (APA) to accredit counseling as well as clinical psychology programs to ensure the training of competent psychologists for the VA system. The VA also was instrumental in encouraging the development of university-based counseling centers to assist veterans with educational and work-related adjustment issues (Meara & Myers, 1999). For these reasons, counseling psychology has historical roots and expertise in career and vocational counseling. Assessments in these areas consider individual differences in career development needs, interests, and barriers to career or employment (Armstrong & Rounds, 2008; Whiston & Rahardja, 2008). Counseling psychologists are in a unique position to address the mental health, educational, and career-planning needs of military veterans and their families because of this historical role and the number of counseling psychologists in college and university settings (Danish & Antonides, 2009).

Currently, one of the primary distinctions between clinical and counseling psychology is the historical focus in clinical psychology on research and practice in the assessment, diagnosis, and treatment of clients with significant psychopathology and emotional disorders. Forensic psychology developed as a subdiscipline within clinical psychology. Although the provision of legal testimony by psychologists dates back to the 1900s, it was not until 2001 that the APA formally recognized forensic psychology as a distinct psychological specialty (Ogloff & Douglas, 2003). In comparison, counseling psychology historically has focused on leveraging and maximizing psychosocial functioning and strengths in individuals who are not experiencing significant psychopathology but are experiencing transitional life stressors (Meara & Myers, 1999).

Thus, the development of clinical and counseling psychology initially was based on the needs of distinct populations. Over time, each discipline has expanded in scope, and each has contributed to assessment process research and practice on the basis of the respective specialty's history and strengths.

THE PURPOSE OF THE PSYCHOLOGICAL ASSESSMENT

The purpose of a psychological assessment is to answer particular questions related to an individual's intellectual, psychological, emotional–behavioral, or psychosocial functioning, or some combination of these domains. These questions are determined by the assessment context and referral source. As Fernandez-Ballesteross (1997) described, a psychological assessment typically is driven by a particular problem or referral question. A psychological *assessment* includes more than psychological testing. Historically, the purpose of a psychological assessment has been to gather information directly from the client, obtain collateral information, administer psychological test instruments, interpret the test results, and provide a conceptualization of the client that integrates the test data with the collateral and interview data. This conceptualization is summarized, a diagnosis or diagnostic rule-out is offered (as applicable), and recommendations are made for consideration related to decision-making (e.g., in career- or education-related choices, personnel decision-making, or parental capacity assessments) and, where appropriate, for treatment. In contrast, psychological *testing* is one component of a psychological assessment. It is measurement oriented. The purpose of testing is to provide a standardized administration of an instrument that has research evidence substantiating the reliability of its scores and the validity of these scores in identifying, quantifying, and describing particular characteristics or abilities when used with a specified population within a specified context. These test scores are interpreted within the context of the client's history and the additional data gathered as part of the assessment process.

THE ASSESSMENT PROCESS

Weiner (2003) described the assessment process as consisting of three phases: information input,

information evaluation, and information output. Each is described here.

Information Input

Information input is the collection of information. It is influenced by the assessment context, referral questions, and referral source. These factors inform why the assessment is requested and what questions are expected to be answered. Such a contextual assessment considers the client's culture and language proficiency when selecting instruments and interpreting instrument scales (Butcher, Cabiya, Lucio, & Garrido, 2007). The referral source and assessment context also influence which instruments are appropriate for use. For example, some instruments appropriate for personality assessment in an outpatient counseling or clinical setting have been found to be inappropriate in a forensic setting because of compromised validity (Carr, Moretti, & Cue, 2005). Selecting appropriate instruments, on the basis of the client's cultural context and the referral context, is the first step in ensuring that the assessment provides valid results for answering the particular referral questions for that particular individual (e.g., Perlin & McClain, 2009).

The Assessment Context and Referral Questions

The referral questions addressed by the assessment are determined by the assessment context. The assessment context also determines the potential sources of collateral information. In turn, the context and referral source determine what requisite education, training, and supervised experience are necessary to conduct the assessment as well as which additional professional standards and guidelines for specialized practice might be applicable.

The assessment context and referral source represent key factors in determining which formal instruments are appropriate, on the basis of the normative sample and ability to identify response patterns. For example, the Millon Clinical Multiaxial Inventory (MCMI; Millon, 1977) was normed and standardized on clients engaged in mental health services. It was not normed on a general population standardization sample (Butcher, 2009). The test developers subsequently reported that the third edition of the MCMI (MCMI-III; Millon, 1994) later was normed on a large sample of newly incarcerated prison inmates for the purpose of predicting adjustment to prison and treatment needs while incarcerated. However, the use of the MCMI-III with populations outside of these standardization samples and for other purposes would be questionable (Butcher, 2009). For further discussion of self-report inventories (and the MCMI-III in particular), readers are referred to Chapter 11, this volume.

Conducting assessments consistent with professional standards and guidelines necessitates staying current with the relevant research. For example, Carr et al. (2005) reported that the Personality Assessment Inventory (PAI; Morey, 1996, as reported in Carr et al., 2005) failed to detect positive self-presentation bias adequately in a sample of 164 parents completing capacity evaluations. This finding suggests that caution should be used in considering the PAI for this type of assessment. However, Boccaccini, Murrie, and Duncan (2006) reported that the PAI Negative Impression Management scale performed as well as the comparison scale (Minnesota Multiphasic Personality Inventory—2 [MMPI–2] F scale) in screening for malingering in a sample of defendants undergoing pretrial evaluations in federal criminal court. Although cross-validation of the results of both studies is important for verifying these conclusions, they underscore the point that an instrument may be appropriate for addressing the referral question in one population yet not perform adequately when the referral question changes and the population differs. Thus, psychologists must pay particular attention to the specific population characteristics, context, and referral questions when selecting test instruments.

Standards and guidelines specific to the type of assessment required and population assessed provide guidance for the selection of appropriate instruments. For example, the APA's *Guidelines for Psychological Practice with Older Adults* (2003) recommend an interdisciplinary approach to the assessment of psychological functioning in older adults. Such an approach facilitates consideration of medication effects and medical conditions on cognitive and emotional functioning. Additional assessment considerations pertinent to this population include

behavioral analyses to identify potential inappropriate or harmful behaviors and interventions to address these behaviors, and a repeated-measures approach to distinguish between stable cognitive and emotional characteristics versus characteristics that are temporally or situation dependent.

The APA (2009) also has issued guidelines for child custody evaluations. A custody evaluation is requested most often when the dissolution of the partner relationship is contentious. What is significant about these evaluations is that the parental assessment is from the perspective of the best psychological interests of the child. The psychologist's role is to provide an impartial opinion that addresses the ability of the parent to provide caretaking consistent with the child's best interests. This task requires that professional opinions or recommendations are based on sufficient objective data to support the psychologist's conclusions (Martindale & Gould, 2007). The assessment assists the court in decision-making concerning the parent's role regarding the physical care, access to, and legal decision-making for the child (APA, 2009).

Parental capacity assessments often are requested in juvenile dependency cases to determine whether a parent's mental health concerns are so severe and incapacitating that the parent cannot safely parent the child or the parent is unable to benefit from services to mitigate the risk of future abuse or neglect of the child. Such assessments require not only requisite education, training, and experience in assessing serious mental illness, including character pathology, but an understanding of judicial and administrative regulations and timelines. Relevant guidelines include the *Guidelines for Psychological Evaluations in Child Protection Matters* (APA, 2011) and the *Specialty Guidelines for Forensic Psychology* (APA, in press). Additional information concerning legal issues in clinical and counseling testing and assessment is provided in Chapter 6, this volume.

Information Evaluation

Information evaluation refers to the interpretation of the assessment data (Weiner, 2003). Accurate interpretation of testing data requires that the psychologist interpret instrument responses and scores according to the test developer's instructions. The general standards and guidelines applicable to conducting psychological assessments across settings and the interpretation of test data include the *Standards for Educational and Psychological Testing* (American Educational Research Association, APA, & National Council on Measurement in Education, 1999, currently under revision) and the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010). The psychologist should consult additional relevant professional standards and guidelines on the basis of the referral source, assessment context, and client characteristics.

An evaluation of the assessment data involves more than scoring and interpreting the instruments administered during the data collection phase of the assessment. The evaluation of assessment data requires a critical evaluation and synthesis of the testing data with the collateral data within the context of the specific referral: the reason for the assessment, the referral source, and referral questions (APA, 2010). Ideally, the psychological assessment utilizes a multidimensional, multisource approach (Allen, 2002; Lachar, 2003) consistent with the multitrait–multimethod matrix developed for construct validation by Campbell and Fiske (1959). A multidimensional, multisource approach entails obtaining formal collateral data by persons close to the client (e.g., family, teacher, probation officer, protective services worker) by means of interview, records, or standardized instruments. Mental health records, school report cards, court reports, and criminal history logs are examples of collateral records. The clinical interview of the client and behavioral observations during the assessment process are additional important sources of data. All of these data provide both convergent and divergent data that can be integrated, synthesized, and summarized to address the referral question. Disconfirming data are particularly useful for guarding against the influence of bias and in assisting in the development of an objective conceptualization of the client (Meyer et al., 2001).

The clinical interview. The client in interview is a central component of the psychological assessment. An unstructured clinical interview allows the psychologist to obtain psychosocial history, psychiatric symptomatology, and the perceived rationale for

the assessment from the client's perspective. These data reflect the client's particular perspective and can be compared with test data and collateral information to assess consistency or divergence across data sources. However, if collateral data are scant or missing, an unstructured interview loses the value of reflecting the client's perspective as clinically relevant information. The unstructured interview may not query symptomatology in a systematic manner. Structured and semistructured interview formats typically include critical diagnostic criteria to facilitate differential diagnosis. Client symptoms are assessed and scores are compared against normative data. However, semistructured and structured interviews still rely on client self-report without the ability to assess response style and test-taking attitude. Thus, all three interview formats are subject to distortion and response bias (Bagby, Wild, & Turner, 2003). Because of this shortcoming, inclusion of formal testing is recommended for inclusion in psychological assessments.

Behavioral observations. Another potential important source of information is the psychologist's careful description of client behavior, test-taking attitude, interactive style, and any special needs that necessitate accommodation or modification of the assessment process or standardized testing procedure. As Leichtman (2009) noted, these behavioral observations can be a rich source of data. In spite of this possibility, Leichtman noted that the behavioral observations section of most assessment reports typically consists of just a few sentences, and training in behavioral observation and reporting tends to be given only superficial treatment in graduate training and supervision. Additionally, despite its descriptive name, the reporting of behavioral observations is prone to subjectivity and bias, another reason why this assessment component warrants careful attention in training as well as self-monitoring by the psychologist during the assessment process (Leichtman, 2009). The psychologist's interpretation and documentation of client behaviors as well as interactive style can be influenced in several ways, such as lack of knowledge or misapplication of base rates for that population and the level of training and competence in assessing clients from that particular population. These topics are discussed in more detail in the section General Assessment Considerations.

Information Output

Information output refers to the utilization of the assessment data to derive conclusions and recommendations that address the referral questions (Weiner, 2003). Accurately synthesizing these data is a complex process that requires critical thinking skills; knowledge of psychological principles, guidelines, and standards related to testing and working with diverse populations; and competence in developing an effective working alliance. These critical thinking skills include an awareness of the relative weight to give to clinical judgment versus actuarial or statistical prediction rules in formulating one's conclusions and guarding against various types of bias in the interpretation and reporting of assessment data.

GENERAL ASSESSMENT CONSIDERATIONS

There are general considerations that apply to all three phases of the assessment process (information input, information evaluation, and information output). For this reason, awareness of these issues guides an appropriate, objective assessment of the client and mitigates the potential for inaccuracy in assessment, synthesis, reporting, and recommendations. These issues include the potential for the introduction of bias and moderator and mediator variables that may influence the working alliance or assessment validity. These two issues may affect any or all of the three phases of the assessment process.

Bias

Test popularity may be considered a type of bias because common usage perpetuates the mistaken belief that an instrument is valid and reliable. For example, the Thematic Apperception Test and other projective techniques are used frequently in clinical and forensic settings, although their use has been seriously questioned (Hunsley & Mash, 2007). An exception may be the Rorschach inkblot method, which has received research support regarding test protocol validity when compared with MMPI protocols (Hiller, Rosenthal, Bornstein, Berry, &

Brunell-Neulieb, 1999). The use of the Rorschach in clinical and forensic settings also has been endorsed by the Society for Personality Assessment (SPA). A thoughtful review of the relevant literature and discussion of the appropriate uses of the Rorschach can be found in the 2005 SPA position statement.

Psychologists also should be aware of the potential for *confirmatory bias*, in which one selectively attends to behaviors that are consistent with the psychologist's expectations or theoretical orientation. These assumptions may be based on the client's cultural or clinical group membership (Sandoval, 1998). A closely related phenomenon is the *availability bias*, in which recent behavior or extreme, vivid behavior is weighted more heavily and is more influential than is warranted by its frequency or clinical significance. These biases result in overinterpretation of assessment data and the potential for overpathologizing the client's behavior or presentation. Seeking out and evaluating sources of potential divergent, as well as convergent (confirmatory), data during the assessment process assists in guarding against confirmatory and availability biases.

Theoretical orientation. The practitioner's theoretical orientation influences the assessment process in terms of instrument selection, questions asked during the clinical interview, and interpretation of client responses and assessment data (Craig, 2009). For these reasons, the psychologist is encouraged to consider the potential for bias. This potential is particularly salient if the psychologist has a background in counseling or clinical mental health and decides to develop competence in completing parental capacity or forensic risk assessments. Theoretical orientation may guide the selection of particular instruments (Lambert & Lambert, 1999). Theoretical orientation or adherence to a particular clinical model also may influence the psychologist's interpretation of test results, resulting in interpretive error regarding diagnosis, etiology, or treatment recommendations. Such errors were first described by Rosenthal (1966) and constitute a phenomenon distinct from experimenter expectancy because they do not influence the client's behavior. This phenomenon also is distinct from test bias because score differences may be statistically and clinically significant (Reynolds & Ramsay, 2003). However, this phenomenon may be associated with (a) the failure to consider relevant base rates (Weiner, 2003); (b) *environmental impressions*, a bias that is based on the particular assessment environment within which the psychologist works (Weiner, 2003); or (c) failure to consider the client's social context, environment, and person–environment interaction (Wright, Lindgren, & Zakriski, 2001).

Base rates. Base rate refers to the actuarial probability that a particular clinical phenomenon, such as a particular diagnosis, will be present in a particular population or assessment context. For example, psychotic disorders are more prevalent in acute inpatient psychiatric settings than in student counseling centers. Bias is introduced when the psychologist inadvertently, or consciously, erroneously applies a base rate probability and fails to consider competing hypotheses or fails to conduct an appropriate differential diagnosis when evaluating assessment data (Weiner, 2003). Understanding the base rates within a particular population also provides a context for evaluating the sensitivity and specificity—and, hence, clinical utility and predictive power—of a particular instrument (Faust, Grimm, Ahern, & Sokolik, 2010).

Assessment of diverse populations. The validity of assessment results generally and test scores in particular may be attenuated when instruments are used inappropriately cross-culturally. In addition to culture, ethnicity, and race, variables known to influence test results and thereby warranting consideration when selecting instruments, include client's primary language, socioeconomic status, and level of education (Gray-Little & Kaplan, 1998).

A starting point in developing cross-cultural competence may be a self-assessment of one's own cultural membership(s). Hays (2008) articulated a clear and structured process for this self-evaluation, which can serve to identify potential biases as a first step in the development of cross-cultural competencies. Migration or immigration history, level of acculturation, and acculturative stress are just three areas of knowledge with which the psychologist should be familiar (Acevedo-Polakovich et al., 2007). When working with culturally diverse

clients, it is important for psychologists to be aware of an instrument's *conceptual equivalence*—that is, the test's ability to measure the same construct across cultures in order to determine its validity for use with a particular client population (Geisinger, 2003). This ability can be determined by comparing evidence of construct validity collected in the "host" language and culture with evidence of construct validity collected in additional linguistic and cultural populations (Geisinger, 2003). Because psychological assessments go beyond test administration and interpretation, Acevedo-Polakovich et al. (2007) suggested "proactive steps" related to initial training that were first offered by Hansen (2002, as reported in Acevedo-Polakovich et al., 2007). These suggestions were specific to the Latina/o population but reflect general principles that could be applied to working with other populations. They include the need to (a) develop an understanding of Latina/o-specific cultural variables, constructs, and syndromes to promote accurate assessment and mitigate the potential for misinterpreting culture-specific beliefs or behaviors; (b) be familiar with instruments of known, and acceptable, validity and reliability with U.S. Latina/os; (c) interpret tests and complete assessments that are consistent with, and relevant to, Latina/o culture; and (d) provide test feedback in a language and style that meet the needs of the client.

The client's personal history and context also influence decision-making regarding the direction of the clinical interview, types of collateral information to collect, and appropriate testing (Comas-Diaz & Grenier, 1998). For example, assessing newcomers (refugees and asylum seekers) includes a careful but nonthreatening querying of where the client came from, when the client left his or her country of origin, and what was going on in that country at that time. The responses to these questions provide a context within which to evaluate the probability that the client experienced torture and consequent mental health symptomatology (Maltzman, 2004).

Sandoval (1998) made the following recommendations to facilitate critical thinking and to guard against bias, particularly when assessing clients from diverse populations: (a) Identify one's own preconceptions in advance to better guard against their influence, (b) ensure that conclusions are drawn after careful consideration, (c) seek appropriate cultural consultation to prevent the misinterpretation of normal behaviors, and (d) ensure that careful notes are taken to prevent reliance on memory.

Moderator and Mediator Variables

Moderator and mediator variables may influence the assessment process in a manner similar to the effects seen in counseling and psychotherapy. Moderator variables include client and psychologist expectations and attitudes about the assessment process. Mediator variables include the behaviors (covert and overt) and client–psychologist interaction that occur during the assessment (Hill & Williams, 2000). Both moderator (input) and mediator (process) variables influence the development of rapport and thus can influence the assessment process and the validity of the collected data and data interpretation.

Developing and maintaining rapport and an effective working alliance is critical to facilitating the assessment process. Despite this necessity, the psychologist has limited time within which to establish a working relationship with the client that promotes cooperation, motivation, and forthrightness in the assessment process.

Client factors. The client's affective state can influence testing and self-report. Anxiety or fear about the testing process may negatively affect attention and concentration and may contribute to mistakes and accidental random responding. In their description of obstacles to establishing rapport from the client's perspective, Lerner and Lerner (1998) described Schafer's (1954, as cited in Lerner & Lerner, 1998) observation that the assessment process requires the client to cede control over what information to hold private and allows intrusiveness by the psychologist without the establishment of a requisite level of trust. The assessment context also can influence the client's approach to participating in the assessment. For example, clients may attempt to minimize symptoms to facilitate discharge from the hospital (Bagby et al., 1997) or present with a defensive style in forensic settings, such as parental custody evaluations (e.g., Bagby, Nicholson, Buis,

Radovanovic, & Fidler, 1999). Traumatized clients may experience the assessment as inherently stressful. They may minimize or deny symptoms in an attempt to avoid remembering and discussing the traumatic events, resulting in the denial of symptoms during the clinical interview and suppressed test scores (Briere, 2004).

Psychologist factors. The psychologist is challenged to engage the client quickly and effectively to promote a collaborative, nondefensive style. In forensic settings, this goal may be difficult to achieve because of the investigative nature of forensic assessments that necessitates a probing, neutral stance in comparison with a more supportive, collaborative role appropriate for a clinical setting (Craig, 2009). In clinical contexts, development of a collaborative working alliance may be impeded if the psychologist is perceived as too distant or inappropriately sympathetic (Briere, 2004). Creed and Kendall (2005) identified therapist variables associated with a positive alliance in therapy with children. These factors included a collaborative stance (in which the therapist encouraged child involvement), not pushing the child to talk when the child was not ready to do so, and emphasizing common ground. Although these variables predicted child ratings of the strength of the therapeutic relationship early in therapy, they did not predict therapist ratings (Creed & Kendall, 2005). This finding suggests that therapists may not be sufficiently sensitive to client responses and reactions in therapy that may mediate the working relationship. These same variables and processes also may be present in and affect the assessment process with children and youths.

As noted earlier, allowing insufficient time to develop a collaborative working relationship and pushing prematurely or inappropriately for information are two psychologist-related variables that may negatively affect the assessment process. Perhaps these behaviors are due, at least in part, to the pressure that psychologists feel to obtain the necessary and sufficient data to answer the referral questions (Lerner & Lerner, 1998). This pressure may feel more acute when the assessment is initiated by a third-party referral who is the payer and the assessment is time sensitive.

Clinical Judgment, Actuarial Prediction, and Utilization of Empirical Guidelines

Meehl's 1954 monograph was the first description of the equivalence or superiority of actuarial prediction in comparison with clinical judgment. Garb (2003) described actuarial prediction as decision rules that are based on empirical data. Actuarial prediction is equivalent to statistical prediction when the latter refers to mathematical equations that are based on empirical data (Garb, 2003). The superiority of actuarial prediction has been confirmed consistently in research, particularly in forensic settings (Ægisdóttir et al., 2006; Garb, 2003). Applied to the assessment process, actuarial prediction is consistent with the utilization of empirical guidelines for deriving assessment conclusions. Weiner (2003) described empirical guidelines as the utilization of decision rules "derive[d] from the replicated results of methodologically sound research" (p. 12). Applying these rules facilitates objective decision-making and mitigates the potential for biases. Empirical guidelines, including the application of appropriate cutoff scores applied within the particular referral context, also mitigate the potential for false-positive or false-negative conclusions (Weiner, 2003). The adoption of an empirical approach also assists in guarding against the influence of confirmatory and personal biases in clinical and counseling settings (Garb, 2003; Heilbrun, DeMatteo, Marczyk, & Goldstein, 2008; Strohmer & Arm, 2006). Despite these findings, psychologists have tended to resist adoption of an empirical approach to assessment and diagnosis (Graham & Naglieri, 2003).

This perceived resistance has been attributed to two primary considerations that reflect an apparent scientist–practitioner split: (a) the need to ensure the construct validity of clinical diagnoses in clinical research versus the time and resource limitations encountered by the clinician in practice and (b) the suboptimal utility of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*) to facilitate treatment planning versus the paramount need for clinical utility of an assessment in treatment settings (Mullins-Sweatt & Widiger, 2009). What do not appear to have consistent support in the literature are the hypotheses that practitioners are

reluctant to adopt empirically derived assessment practices because of philosophical differences or that practitioners believe that empirically derived diagnoses are simplistic or invalid (Widiger & Samuel, 2009).

Conversely, researchers have acknowledged that the psychological assessment of an adult or child in a clinical mental health setting must address diagnostic clarification for the purpose of treatment planning, prediction of response to treatment, and prognosis for future level of functioning (Bagby et al., 2003; Lachar, 2003). In other words, the clinical utility of the assessment is paramount (Mullins-Sweatt & Widiger, 2009). Despite the research and general consensus supporting the superiority of empirically based assessment, formal psychological testing and structured or semistructured interviews are not always utilized in clinical practice (Widiger & Samuel, 2009). Failure to use standardized assessment procedures potentially compromises the validity and reliability of the resulting clinical diagnoses. This possibility is magnified if, as reported, clinicians do not consistently and routinely adhere to *DSM–IV* diagnostic criteria when utilizing an unstructured interview format (Mullins-Sweatt & Widiger, 2009; Widiger & Samuel, 2009). Such lapses may occur because the client's self-report may not be candid or because the clinician may not adequately query the client. For this reason, there is an increased risk that the assessment will be compromised, resulting in a diagnosis (or diagnoses) that does not fully describe the client's presentation and functioning. The resulting diagnoses may, in turn, result in inappropriate or inadequate treatment. In particular, failure to assess for the presence of personality disorder or maladaptive personality traits may compromise not only appropriate treatment but also the accuracy of the predicted response to treatment and posttreatment prognosis (Widiger & Samuel, 2009). Widiger and Samuel suggested a tiered approach to the assessment of personality disorder to bridge this schism. The initial tier would be administration of a self-report inventory, such as the MMPI–2–RF (Restructured Form) (Ben-Porath & Tellegen, 2008) or the MCMI-III (Millon, 1994), which would be followed by a semistructured interview targeting personality traits identified as maladaptive through the self-report inventory. The goal of this tiered approach is to shorten the semistructured interview to target more carefully the personality traits that appear most salient, thus saving the practitioner time. Whether this approach is disseminated and adopted within the practice community remains to be seen. However, a potential obstacle to this approach may be the reluctance of third-party payers to reimburse for any testing or low reimbursement rates when testing is authorized.

Therefore, rather than a philosophical reluctance, it may be that reimbursement and resource issues are primary factors contributing to practitioners' reluctance to implement empirical assessment approaches.

EMERGING TRENDS

Multiple factors, including the mental health consumer movement, government oversight, and reimbursement policies of third-party payers, have contributed to the call for psychology to demonstrate that its services are cost effective, are measurable, and benefit clients in tangible ways. Three emerging trends in assessment are particularly salient within this context: assessing psychosocial functioning, assessing outcomes, and utilizing the assessment as treatment.

Assessment of Psychosocial Functioning

Over the past 20 years, there has been increasing emphasis within clinical settings to assess the client's psychosocial functioning in addition to psychiatric symptomatology. Psychosocial functioning includes assessment of the client's hobbies, leisure activities, and pursuit of values that are hypothesized to contribute to psychological and subjective well-being (Robbins & Kliewer, 2000). Thus, psychosocial functioning as a construct is expanded to include the assessment of self-enhancing activities in addition to traditional areas of basic functioning such as activities of daily living, interpersonal relationships, and participation in work or school. This conceptualization of psychosocial functioning more clearly articulates the assessment of client strengths in addition to deficits. This strengths-based

approach is the result of several converging areas of research and public policy, including the following:

- the mental health consumer movement (e.g., Campbell & Leaver, 2003; Pulice & Miccio, 2006),
- the rise of the biopsychosocial model in psychology (e.g., Maltzman, 2012), and
- developmental research in the physiological and psychosocial bases of resilience (e.g., Greenberg, 2006; Werner, 2005).

Ro and Clark (2009) described their initial efforts to clarify the construct of psychosocial functioning. The goal of the factor analysis was to initiate the development of a psychometrically sound instrument that could be used to assess the psychosocial deficits associated with *DSM* Axis I and Axis II psychopathology. A community sample ($N = 429$) that included almost equivalent numbers of students and nonstudent residents completed measures assessing quality of life, daily functioning, and personality functioning. Two principal-axis factor analyses with promax rotation were conducted that included measures of functioning across a variety of domains and with varying levels of specificity and breadth. The first factor analysis excluded the two measures of personality functioning, the Measure of Disordered Personality and Functioning (MDPF; Parker et al., 2004, as cited in Ro & Clark, 2009) and the Severity Indices of Personality Problems (SIPP; Verheul et al., 2008, as cited in Ro & Clark, 2009). By excluding and then including these measures, these investigators were able to explore whether personality functioning, as defined by these instruments, improved the factor solution. A four-factor solution, which included these personality functioning measures, yielded the most psychologically interpretable solution (Ro & Clark, 2009). The resulting four dimensions reflected Basic Functioning (activities of daily living and microlevel functioning), Well-Being (subjective sense of well-being, satisfaction, and high social functioning), and two factors on which the MDPF and SIPP loaded: Self-Mastery (impulsivity, inability to learn from experience, and lack of self-control) and Interpersonal and Social Relationships (lack of empathy or caring for others, difficulty fitting in socially). These two personality functioning measures were interpreted by these investigators as reflecting social and environmental functioning associated with personality traits. Ro and Clark noted that they could only include general measures of psychosocial functioning that were applicable across a range of client populations.

The growing emphasis on psychosocial functioning reflects the growing imperative to demonstrate the clinical utility of the assessment, defined as the ability to demonstrate that the assessment "makes a difference with respect to the accuracy, outcome, or efficiency of clinical activities" (Hunsley & Mash, 2007, p. 45). This imperative has been an impetus for developing assessment instruments with adequate external validity to ensure that assessment results reflect the client's capacity to function in "real-world" settings (Kubiszyn et al., 2000). Neuropsychologists have acknowledged this need as their field has shifted from an emphasis on descriptive diagnosis toward clarifying functional capacity and recommending specific rehabilitative interventions (Rabin, Burton, & Barr, 2007). In particular, there is increased emphasis in ensuring instrument ecological validity defined as the generalizability of test results assessed in a controlled setting to the actual skill sets required in daily living (Rabin et al., 2007). A potential advantage of developing and utilizing ecologically oriented instruments (EOIs) is that they could minimize the potential for the misinterpretation of test scores on the basis of client variables known to influence neuropsychological test results.

The confluence of three factors—(a) the growing emphasis on psychosocial functioning, (b) the emergence of EOIs in neuropsychology, and (c) the acknowledgment of the superiority of actuarial and evidence-based assessment measures—may provide the impetus to look beyond self-report instruments in clinical psychology toward the development of more ecologically valid assessments of psychological functioning.

Assessment as Treatment

As noted earlier in this chapter, the assessment context as well as psychologist-related and client-related variables can influence the establishment of rapport and the working alliance. In clinical settings, the

psychological assessment is often the precursor to treatment. One consistent finding in psychotherapy process and outcomes research is that a strong positive working alliance established early in therapy correlates with a decreased probability of early termination and predicts achievement of treatment goals and positive therapy outcomes (Hilsenroth & Cromer, 2007). Extrapolating from these findings, Finn and colleagues (e.g., Finn & Tonsager, 1997) developed the Therapeutic Model of Assessment (TMA), the goal of which is the use the assessment process as a treatment intervention. For a detailed treatment of therapeutic assessment, readers should consult Chapter 26, this volume.

The TMA integrates the multimethod approach to information gathering with an empathic, collaborative approach in which the test feedback session becomes an intervention: "The major goal is for clients to leave their assessments having had new experiences or gained new information about themselves that subsequently helps them make changes in their lives" (Finn & Tonsager, 1997, p. 378). This client-empowering, collaborative, strengths-based approach to clinical assessment is consistent with counseling psychology's historical approach to vocational, career, and personal counseling (Delworth, 1977; Fretz, 1985; Super, 1955). In the TMA, the assessment and, particularly, the test feedback session become the first phase of treatment. Because the TMA facilitates treatment by means of the assessment process, it may be viewed favorably by third-party payers who otherwise might be reluctant to preauthorize and pay for a formal psychological assessment. The TMA assumes that the same psychologist conducts the assessment and provides the therapy. In some clinical settings, this may be appropriate from an ethical perspective (APA, 2010). In other settings and contexts, particularly forensic settings, the provision of assessment and treatment by the same psychologist could be considered a violation of professional standards (APA, 2010, in press).

Assessing Outcomes

With the advent of managed health care and time-limited treatments, there is increased interest on the part of third-party payers for psychologists to demonstrate the clinical utility of the psychological assessment to justify its cost (Hunsley & Mash, 2005). The public sector (i.e., government agencies) and the private sector (behavioral health care insurance companies) have increased the pressure on mental health professionals to demonstrate the effectiveness of their treatments and interventions (e.g., APA Practice Directorate, 2007; Cavaliere, 1995). This pressure is not likely to abate as financial resources dwindle and public scrutiny regarding the expenditure of government money increases. Although these external bodies are cited as the sources of this pressure, psychology as a profession also historically has demanded that services demonstrate effectiveness to justify reimbursement and inclusion in national health care initiatives. These pressures, from outside and within psychology, were a significant impetus for the development of treatment outcomes research (Maltzman, 2012).

Hill and Corbett (1993) defined outcomes as the changes that result, either directly or indirectly, from the treatment utilized in counseling or psychotherapy. Assessment instruments that can monitor progress in treatment as well as address the referral question have fundamental advantages over instruments that can be used as part of the assessment but whose cost, time, length, or other factors preclude their use over the course of treatment. In addition to tracking individual client progress over time, instruments that can be used as a repeated measure for tracking individual progress over time also can facilitate continuous quality improvement efforts at the organizational or system level by aggregating and analyzing data across clients. The assessment of outcomes necessitates a multimodal approach to ensure that the clinically salient variables targeted in treatment are adequately assessed over time and sufficiently sensitive to detect change over time (Lambert & Lambert, 1999).

Historically, the assessment of outcomes has focused on Axis I clinical disorders, which exclude personality disorders and mental retardation (American Psychiatric Association, 2000), and areas of functioning compromised by these disorders. However, development of instruments for the assessment and change over time of personality functioning, such as capacity for empathy and tendency toward impulsivity, would be enormously helpful in mitigating risk

in forensic settings as well as for potentiating treatment of Axis I disorders in clinical and counseling settings (Widiger & Samuel, 2009). For additional discussion of risk assessment in forensic settings, readers are directed to Chapter 16, this volume.

SUMMARY AND DISCUSSION

The assessment process historically has consisted of a multimethod approach integrating interview, collateral, and formal test data. Both clinical and counseling psychology have brought strengths to the process that are based on the historical populations served by each discipline and referral questions addressed. Clinical psychology introduced psychological testing and the multimethod approach for the assessment of emotional disturbance in children. Counseling psychology emerged to address the vocational needs of these youths. Both disciplines transitioned into the assessment of adults with clinical psychology focusing on the assessment and treatment of major psychopathology. Counseling psychology historically has focused on the assessment and treatment of life-associated stressors in individuals functioning along the continuum of normal psychological functioning. Both specialties have strong bases in empiricism and formal psychological testing. Their historical convergence may be in the assessment process itself. One emerging trend is the increasing focus on psychologist–client collaboration during the assessment, which essentially becomes the initiation of treatment (e.g., Tharinger et al., 2009). Counseling psychologists historically have collaborated with clients, together reviewing test data and their application to vocational and career choices (Swanson & Gore, 2000). This approach has naturally segued into personal counseling for adjustment issues. Clinical psychology appears to be adapting this approach to the process of the clinical assessment for psychotherapy. The assessment context and referral questions will determine the extent to which this collaborative approach is appropriate. In most forensic settings, it may be very limited or inconsistent with applicable professional standards and guidelines.

Balancing the use of subjective sources of data (e.g., the clinical interview and most self-report instruments) and objective sources of data (e.g., behavioral analyses, test instruments with validity scales) is a topic of continuing discussion and varied practice. Ensuring that multiple methods are used for data collection helps guard against the introduction of biases that can occur if subjective data sources predominate. Adherence to professional standards and guidelines, education and training in assessing diverse populations, and awareness of various sources of bias also facilitate an assessment process that results in a data synthesis and report that can objectively address the referral questions.

References

Acevedo-Polakovich, I. D., Reynaga-Abika, G., Garriott, P. O., Derefinko, K. J., Wimsatt, M. K., Gudonis, L. C., & Brown, T. L. (2007). Beyond instrument selection: Cultural considerations in the psychological assessment of U.S. Latinas/os. *Professional Psychology: Research and Practice, 38*, 375–384. doi:10.1037/0735-7028.38.4.375

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382. doi:10.1177/0011000005285875

Allen, J. B. (2002). *Treating patients with neuropsychological disorders: A clinician's guide to assessment and referral*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (2003). *Guidelines for psychological practice with older adults*. Washington, DC: Author. Retrieved from http://www.apa.org/practice/guidelines/older-adults.pdf

American Psychological Association. (2009). *Guidelines for child custody evaluations in family law proceedings*. Washington, DC: Author. Retrieved from http://www.apa.org/practice/guidelines/child-custody.pdf

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, amended June 1, 2010)*. Washington, DC: Author.

Retrieved from http://www.apa.org/ethics/code/index.aspx

American Psychological Association. (2011). *Guidelines for psychological evaluations in child protection matters*. Washington, DC: Author. Retrieved from http://www.apa.org/practice/guidelines/child-protection.pdf

American Psychological Association. (in press). *Specialty guidelines for forensic psychology*. Washington, DC: Author.

American Psychological Association Practice Directorate. (2007). APA group to propose pay-for-performance policy. *Monitor on Psychology, 38*(4), 33. Retrieved from http://psycnet.apa.org/psycextra/599732007-024.pdf

Armstrong, P. I., & Rounds, J. B. (2008). Vocational psychology and individual differences. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (4th ed., pp. 375–391). Hoboken, NJ: Wiley.

Bagby, R. M., Nicholson, R. A., Buis, T., Radovanovic, H., & Fidler, B. J. (1999). Defensive responding on the MMPI-2 in family custody and access evaluation. *Psychological Assessment*, *11*, 24–28. doi:10.1037/1040-3590.11.1.24

Bagby, R. M., Rogers, R., Nicholson, R. A., Buis, T., Seeman, M. V., & Rector, N. A. (1997). Effectiveness of MMPI–2 validity indicators in the detection of defensive responding in clinical and nonclinical samples. *Psychological Assessment*, *9*, 406–413. doi:10.1037/1040-3590.9.4.406

Bagby, R. M., Wild, N., & Turner, A. (2003). Psychological assessment in adult mental health settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 213–234). Hoboken, NJ: Wiley.

Baker, D. B. (1988). The psychology of Lightner Witmer. *Professional School Psychology, 3*, 109–121. doi:10.1037/h0090552

Baker, D. B. (2002). Child saving and the emergence of vocational psychology [Abstract]. *Journal of Vocational Behavior, 60*, 374–381. doi:10.1006/jvbe.2001.1837

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI–2–RF: Manual for administration, scoring, and interpretation.* Minneapolis, MN: University of Minnesota Press.

Boccaccini, M. T., Murrie, D. C., & Duncan, S. A. (2006). Screening for malingering in a criminal-forensic sample with the Personality Assessment Inventory. *Psychological Assessment, 18*, 415–423. doi:10.1037/1040-3590.18.4.415

Briere, J. (2004). *Psychological assessment of adult posttraumatic states: Phenomenology, diagnosis, and measurement* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/10809-000

Butcher, J. N. (2009). Overview and future directions. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 707–718). New York, NY: Oxford University Press.

Butcher, J. N., Cabiya, J., Lucio, E., & Garrido, M. (Eds.). (2007). The challenge of assessing clients with different cultural and language backgrounds. In *Assessing Hispanic clients using the MMPI–2 and MMPI–A* (pp. 3–23). Washington, DC: American Psychological Association. doi:10.1037/11585-001

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Campbell, J., & Leaver, J. (2003). *Emerging new practices in organized peer support.* Retrieved from http://www.nasmhpd.org/nasmhpd_collections/collection5/publications/ntac_pubs/reports/peer%20support%20practices%20final.pdf

Carr, G. D., Moretti, M. M., & Cue, B. J. H. (2005). Evaluating parenting capacity: Validity problems with the MMPI–2, PAI, CAPI, and ratings of child adjustment. *Professional Psychology: Research and Practice, 36*, 188–196. doi:10.1037/0735-7028.36.2.188

Cavaliere, F. (1995). Measuring outcomes: Payers demand increased provider documentation. *APA Monitor*, *26*(10), 41.

Comas-Díaz, L., & Grenier, J. R. (1998). Migration and acculturation. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 213–239). Washington, DC: American Psychological Association. doi:10.1037/10279-008

Craig, R. J. (2009). The clinical interview. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 201–225). New York, NY: Oxford University Press.

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. doi:10.1037/0022-006X.73.3.498

Danish, S. J., & Antonides, B. J. (2009). What counseling psychologists can do to help returning veterans. *The Counseling Psychologist, 37*, 1076–1089. doi:10.1177/0011000009338303

Delworth, U. (1977). Counseling psychology. *The Counseling Psychologist, 7*, 43–45. doi:10.1177/001100007700700219

Faust, D., Grimm, P. W., Ahern, D. C., & Sokolik, M. (2010). The admissibility of behavioral science evidence in the courtroom: The translation of legal

to scientific concepts and back. *Annual Review of Clinical Psychology, 6*, 49–77.

Fernandez-Ballesteross, R. (1997). Guidelines for the assessment process. *European Psychologist, 2*, 352–355. doi:10.1027/1016-9040.2.4.352

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9*, 374–385. doi:10.1037/1040-3590.9.4.374

Fretz, B. R. (1985). Counseling psychology. In E. M. Altmaier & M. E. Meyer (Eds.), *Applied specialties in psychology* (pp. 45–73). New York, NY: Random House.

Garb, H. N. (2003). Clinical judgment and mechanical prediction. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 27–42). Hoboken, NJ: Wiley.

Geisinger, K. F. (2003). Testing and assessment in cross-cultural psychology. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 95–117). Hoboken, NJ: Wiley.

Graham, J. R., & Naglieri, J. A. (2003). Current status and future directions of assessment psychology. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 579–592). Hoboken, NJ: Wiley.

Gray-Little, B., & Kaplan, D. A. (1998). Interpretation of psychological tests in clinical and forensic evaluations. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 141–178). Washington, DC: American Psychological Association. doi:10.1037/10279-006

Greenberg, M. T. (2006). Promoting resilience in children and youth: Preventive interventions and their interface with neuroscience. In B. M. Lester, A. Masten, & B. McEwen (Eds.), *Annals of the New York Academy of Sciences: Vol. 1094. Resilience in children* (pp. 139–150). New York: New York Academy of Sciences.

Hays, P. A. (2008). Looking into the clinician's mirror: Cultural self-assessment. In P. A. Hays (Ed.), *Addressing cultural complexities in practice: Assessment, diagnosis, and therapy* (2nd ed., pp. 41–62). Washington, DC: American Psychological Association. doi:10.1037/11650-003

Heilbrun, K., DeMatteo, D., Marczyk, G., & Goldstein, A. M. (2008). Standards of practice and care in forensic mental health assessment. *Psychology, Public Policy, and Law, 14*, 1–26. doi:10.1037/1076-8971.14.1.1

Hill, C. E., & Corbett, M. M. (1993). A perspective on the history of process and outcome research in counseling psychology. *Journal of Counseling Psychology, 40*, 3–24. doi:10.1037/0022-0167.40.1.3

Hill, C. E., & Williams, E. N. (2000). The process of individual therapy. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 670–710). New York, NY: Wiley.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neulieb, S. (1999). A comparative analysis of Rorschach and MMPI validity. *Psychological Assessment, 11*, 278–296. doi:10.1037/1040-3590.11.3.278

Hilsenroth, M. J., & Cromer, T. D. (2007). Clinician interventions related to alliance during the initial interview and psychological assessment. *Psychotherapy: Theory, Research, Practice, Training, 44*, 205–218. doi:10.1037/0033-3204.44.2.205

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment, 17*, 251–255. doi:10.1037/1040-3590.17.3.251

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29–51.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., . . . Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice, 32*, 119–130. doi:10.1037MJ73S-702S.31.2.119

Lachar, D. (2003). Psychological assessment in child mental health settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 235–260). Hoboken, NJ: Wiley.

Lambert, M. J., & Lambert, J. M. (1999). Use of psychological tests for assessing treatment outcome. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 115–151). Mahwah, NJ: Erlbaum.

Leichtman, M. (2009). Behavioral observations. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 187–200). New York, NY: Oxford University Press.

Lerner, P. M., & Lerner, H. D. (1998). An experiential psychoanalytic approach to the assessment process. In J. W. Barron (Ed.), *Making diagnosis meaningful: Enhancing evaluation and treatment of psychological disorders* (pp. 247–266). Washington, DC: American Psychological Association. doi:10.1037/10307-009

Maltzman, S. (2004, July). Newcomer women: Co-morbid mental health and physical health concerns. In K. L. Norsworthy (Chair), *Feminist perspectives in international psychology: Building partnerships and creative collaboration.* Roundtable conducted at the 112th Annual Convention of the American Psychological Association, Honolulu, HI.

Maltzman, S. (2012). Process and outcomes in counseling and psychotherapy. In E. M. Altmaier & J. C. Hansen (Eds.), *The Oxford handbook of counseling psychology* (pp. 95–127). New York, NY: Oxford University Press.

Martindale, D. A., & Gould, J. W. (2007). Custody evaluation reports: The case for empirically derived information. *Journal of Forensic Psychology Practice*, *7*, 87–99. doi:10.1300/J158v07n03_06

McGuire, F. L. (1990). *Psychology aweigh! A history of clinical psychology in the United States Navy, 1900–1988*. Washington, DC: American Psychological Association. doi:10.1037/10069-001

Meara, N. M., & Myers, R. A. (1999). A history of Division 17 (Counseling Psychology): Establishing stability amid change. In D. A. Dewsbury (Ed.), *Unification through division: Histories of the divisions of the American Psychological Association* (Vol. 3, pp. 9–41). Washington, DC: American Psychological Association. doi:10.1037/10281-001

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. doi:10.1037/11281-000

Meyer, G. J., Finn, S. D., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Millon, T. (1977). *Millon Clinical Multiaxial Inventory*. Minneapolis, MN: National Computer Systems.

Millon, T. (1994). *Millon Clinical Multiaxial Inventory—III: Manual*. Minneapolis, MN: Pearson Assessments.

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and *DSM–V*. *Psychological Assessment, 21*, 302–312. doi:10.1037/a0016607

Ogloff, J. R., & Douglas, K. S. (2003). Psychological assessment in forensic psychology. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 345–363). Hoboken, NJ: Wiley.

Perlin, M. L., & McClain, V. (2009). "Where souls are forgotten": Cultural competencies, forensic evaluations, and international human rights. *Psychology, Public Policy, and Law, 15*, 257–277. doi:10.1037/a0017233

Pulice, R. T., & Miccio, S. (2006). Patient, client, consumer, survivor: The mental health consumer movement in the United States. In J. Rosenberg & S. Rosenberg (Eds.), *Community mental health: Challenges for the 21st century* (pp. 7–14). New York, NY: Routledge.

Rabin, L. A., Burton, L. A., & Barr, W. B. (2007). Utilization rates of ecologically oriented instruments among clinical neuropsychologists. *The Clinical Neuropsychologist, 21*, 727–743. doi:10.1080/13854040600888776

Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 67–93). Hoboken, NJ: Wiley.

Ro, E., & Clark, L. A. (2009). Psychosocial functioning in the context of diagnosis: Assessment and theoretical issues. *Psychological Assessment, 21*, 313–324. doi:10.1037/a0016607

Robbins, S. B., & Kliewer, W. L. P. (2000). Advances in theory and research on subjective well being. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 310–345). New York, NY: Wiley.

Rosenthal, R. (1966). Interpretation of data. In R. Rosenthal (Ed.), *Experimenter effects in behavioral research* (pp. 16–26). New York, NY: Meredith.

Sandoval, J. (1998). Critical thinking in test interpretation. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 31–49). Washington, DC: American Psychological Association. doi:10.1037/10279-002

Society for Personality Assessment. (2005). The status of the Rorschach in clinical and forensic settings: An official statement by the Board of Trustees of the Society for Personality Assessment. *Journal of Personality Assessment, 85*, 219–237. doi:10.1207/s15327752jpa8502_16

Strohmer, D. C., & Arm, J. R. (2006). The more things change, the more they stay the same: Reaction to Ægisdóttir et al. *The Counseling Psychologist*, *34*, 383–390. doi:10.1177/0011000005285879

Super, D. E. (1955). Transition: From vocational guidance to counseling psychology. *Journal of Counseling Psychology, 2*, 3–9. doi:10.1037/h0041630

Swanson, J., & Gore, P., Jr. (2000). Advances in vocational psychology theory and research. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 233–269). New York, NY: Wiley.

Tharinger, D. J., Finn, S. E., Gentry, L., Hamilton, A., Fowler, J., Matson, M., . . . Walkowiak, J. (2009). Therapeutic assessment with children: A pilot study of treatment acceptability and outcome. *Journal of Personality Assessment*, *91*, 238–244. doi:10.1080/00223890902794275

Weiner, I. B. (2003). The assessment process. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 3–25). Hoboken, NJ: Wiley.

Werner, E. E. (2005). Resilience research: Past, present, and future. In R. DeV. Peters, B. Leadbeater, & R. J. McMahon (Eds.), *Resilience in children, families, and communities: Linking context to practice and policy* (pp. 3–11). New York, NY: Kluwer Academic/ Plenum.

Whiston, S. C., & Rahardja, D. (2008). Vocational counseling process and outcome. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (4th ed., pp. 444–461). Hoboken, NJ: Wiley.

Widiger, T. A., & Samuel, D. B. (2009). Evidence-based assessment of personality disorders. *Personality Disorders: Theory, Research, and Treatment*, *S*(1), 3–17. doi:10.1037/1949-2715.S.1.3

Wright, J. C., Lindgren, K. P., & Zakriski, A. L. (2001). Syndromal versus contextualized personality assessment: Differentiating environmental and dispositional determinants of boys' aggression. *Journal of Personality and Social Psychology, 81*, 1176–1189. doi:10.1037/0022-3514.81.6.1176

CHAPTER 3

COMMUNICATING TEST RESULTS

Virginia Smith Harvey

Effectively communicating test results, both orally and in writing, is challenging because it serves several purposes for multiple audiences. Well-presented oral reports convey test results so clearly and persuasively that the audience understands complex information and is motivated to use suggested changes. Well-written reports convert "assessment data into faithfully designed and executed interventions that lead to improved . . . performance" (Surber, 1995, p. 161); provide legal and historical documentation of the assessment for service eligibility (Ownby, 1997); link data to theory; organize, integrate, and clarify gaps in test data (Blais & Smith, 2008); and enable both the psychologist and consumer to interpret and synthesize results with other ecological and systemic data. Well-communicated test results transform background information, observations, data, test scores, and current contextual variables into hypotheses that evolve into useful descriptions, diagnoses, and appropriate and helpful interventions. Consequently, psychological reports are "one of the most crucial parts of the evaluation process" (Nuttall, Devaney, Malatesta, & Hampel, 1999, p. 396).

The inherent challenge of communicating test results is illustrated by a recent observation during a school-based meeting regarding an eighth grader. The parents, both of whom had doctorates, had read the assessment reports before the meeting. After the clinical psychologist, school psychologist, speech and language pathologist, and special education teacher described their assessment results, the chairperson asked the parents whether they were satisfied with the assessments. After a few minutes of uncomfortable silence, the father explained that they were not interested in filing a formal complaint but that what they wished they had received was an explanation of "what is wrong and what to do about it." The assessing psychologists were shocked, believing that they had already thoroughly addressed those questions both during the meeting's first 90 minutes and in the written reports that the parents had already read. Despite these efforts, however, they had not communicated the test results well enough for even highly educated parents. Such a lack of communication is distressingly common.

Some of this failure to communicate effectively stems from considerable confusion regarding the appropriate approach and focus of psychological reports. As Groth-Marnat (2009b) indicated, psychological reports can be (a) *literary*, using everyday language, creativity, and a dynamic approach that may not be credible; (b) *clinical*, focusing on areas in need of change but disregarding strengths; (c) *scientific*, stressing objectivity and factual information but lacking warmth and positive regard; or (d) *professional*, which balances the strengths of each of the aforementioned and is consumer focused. Consumer-focused assessments accommodate the perspectives of both the referring party and report recipients, clearly and concisely present the data, and include useful and appropriate recommendations in a manner that supports implementation (Brenner, 2003; Groth-Marnat, 2009b; Whitaker, 1994). To write consumer-focused reports, psychologists must determine both the underlying purposes

DOI: 10.1037/14048-003
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

of the assessment and identify the primary consumers of the psychological reports. These factors affect every aspect of the assessment process: tests selected, report content, recommendation selection, conclusions drawn, and language and tone of the ensuing report.

This chapter assumes that underlying purposes of psychological evaluations are to communicate information regarding clients' social, vocational, academic, behavioral, cognitive, and emotional functioning; offer consumers fresh perspective regarding assessed clients; and provide specific and appropriate recommendations that eventually enhance clients' functioning. The presumption is that all of these purposes are best accomplished through a problem-solving orientation that is contextually driven. Finally, it assumes that the most common—and, therefore, primary—individuals to whom test results are reported are nonpsychologists: adult clients, health professionals, education professionals, and other consumers, such as parents.

Psychologists who write psychological reports encounter the aforementioned challenges regardless of their professional domain, and this chapter is therefore intended to be useful to clinical, counseling, industrial, and school psychologists alike. In this spirit, an effort has been made to include a variety of examples from various domains.

ORAL COMMUNICATION OF TEST RESULTS

Usually, test results are communicated orally to consumers as well as in a written report. The traditional sequence is to hold a conference in which the results are orally described and then distribute the written report at a later date. As Ritzler (1998) indicated, generally it is "neither advisable nor ethical to give the report prior to the verbal feedback" (p. 424). During the verbal feedback session, the psychologist gives a general interpretation of results and elicits feedback to ensure that the consumer accurately understands the information. This practice enables the psychologist to modulate the discussion according to the understanding of the audience. Such meetings also enable the psychologist to include clients (and parents, when the assessment involves a child) more completely as collaborative partners in the assessment process. For example, psychologists can verbally share impressions immediately after administering tests; check how well those impressions match those of the client (and parent, when the assessment involves a child); and invite them to add to, clarify, or disagree with information in the report prior to dissemination. In addition to increasing trust, including clients as collaborators in the report-writing process can serve as a therapeutic tool because it results in increased client empowerment, realistic goal setting, and commitment (St. George & Wulff, 1998). All of these practices enable psychologists to obtain data that are not "obfuscated by mistrust, misunderstanding, and the inhibition of self-disclosure" (Bersoff, 1995, p. 286).

Psychologists can use several strategies to facilitate their oral test reporting. First, it is helpful to select primary points to emphasize. As illustrated by the vignette at the beginning of the chapter, these points are likely to be strengths and weaknesses as well as what can be done to improve the situation. These points should be the desired "take-away" message that the consumer will report to friends and relatives. Second, psychologists should accompany oral information with visual aids. These are likely to include graphs depicting progress as well as normal curves that enable psychologists to point out standard and percentile scores relative to the general population. Third, psychologists should encourage recipients to ask questions and give corrective feedback throughout the presentation and should ensure that these corrections are incorporated into the final report. Finally, psychologists should schedule sufficient time for meetings, thereby allowing adequate time for recipients to process the results and reach closure. Although these processes are time consuming, they are requisite for consumers to reach the level of understanding needed before they can appropriately implement recommendations.

Unfortunately, some state departments of education require that parents be sent a copy of a school district report 2 days before the meeting in which the report is reviewed. This practice results in parents reading and attempting to understand psychological reports without assistance and can lead to serious misunderstanding. An alternate sequence for

such situations would be for the psychologist to hold an informal meeting, before the formal meeting, during which the results are interpreted without firm recommendations.

To monitor the effectiveness of their oral communication of test results, psychologists may consider collecting information regarding their effectiveness by asking clients (and parents, when the client is a child) to complete a brief survey revealing their understanding of the material presented and satisfaction with the explanations given. Such satisfaction surveys are being increasingly used in response to medical visits and can serve an important role in leading to improved practice.

WRITTEN COMMUNICATION OF TEST RESULTS

As described by Ownby (2009), psychological report organization varies considerably. "Professional letter" reports are short, contain only a summary of test results, and list prescriptive recommendations. Longer, more traditional psychological reports can be organized in a number of ways. Test-oriented reports are organized according to the tests administered, ability- or domain-oriented reports are organized according to the skills that are assessed on multiple tests (e.g., working memory), and hypothesis-oriented reports are organized to address multiple hypotheses that may explain the client's difficulties (e.g., posttraumatic stress disorder or anxiety).

Many contemporary authors recommend organizing psychological reports by themes rather than test by test (Brinkman, Segool, Pham, & Carlson, 2007; Groth-Marnat, 2009a, 2009b; Nuttall et al., 1999; Ritzler, 1998; Sattler, 2008). In a theme-based report, all the data gathered through file reviews, observations, interviews, self-report scales, informal assessment tools, and standardized tests that relate to a particular topic, such as social anxiety, are grouped together. Similar groupings occur for each theme, determined by referral questions. In addition, the order of discussion is driven by the most important referral question rather than invariably beginning by reporting cognitive assessment results (Ritzler, 1998).

In theme-based reports, psychologists integrate and contextualize the information they have uncovered from a variety of sources. In contrast, in test-by-test reports, this task is implicitly assigned to the consumer. Unfortunately, many consumers do not have the skills to accomplish this integration themselves. Pelco, Ward, Coleman, and Young (2009) found that teachers, particularly less experienced teachers, were better able to understand and develop classroom-targeted interventions after reading reports that were theme based and relatively jargon free. When confronted with test-by-test reports, consumers typically respond by ignoring test interpretations in reports and reading only the brief summary (Surber, 1995). In addition, reports written in a test-by-test format require more validity statements to maintain credibility (Groth-Marnat & Horvath, 2006).

Readability

Consumers should be able to read and understand psychological reports. As Ackerman (2006) indicated, a report's audience may be mental health professionals, nonmental health professionals, or nonprofessionals such as parents. In truth, it is likely to be all three, because psychological reports reach a broad audience both at the time of the assessment itself and for years after. Legislation such as the Family Educational Rights and Privacy Act (1974) and the Individuals with Disabilities Education Improvement Act (IDEA; 2004) established the right of clients (and of parents when the client is a child) to receive copies of psychological and psychoeducational reports. Consequently, reports are commonly distributed to clients (and to parents when the client is a child) in addition to physicians, educators, and other psychologists. Nuttall et al. (1999) concluded that although it may be ideal to write two reports—one to share with fellow psychologists and one to share with nonpsychologists—normally, only one version of a report is written. Therefore, they recommended that psychologists use "plain English." Thus, it is critical that reports be written in a manner that is understandable to a population who are not psychologists and who may be unable to read above the 12th-grade level.

The importance of clarity in communicating test results has been repeatedly stressed by numerous authors (Brenner, 2003; Cuadra & Albaugh, 1956;

Groth-Marnat, 2009b; Groth-Marnat & Horvath, 2006; Harvey, 1997; Kamphaus, 1993; Koocher & Keith-Spiegel, 1998; Martin, 1972; Ownby, 1997, 2009; Sattler, 2008; Tallent, 1993; Whitaker, 1994). Nonetheless, psychological reports are frequently written at a readability level considerably above the education level of their recipients and contain excessive jargon and ill-defined terms (Whitaker, 1994). This lack of clarity persists even though studies have demonstrated that recipients prefer clearly explained technical terms, clear examples, comprehensible explanations, and understandable solutions (Pryzwansky & Hanania, 1986; Rucker, 1967; Shively & Smith, 1969).

For assessment of the readability of written material, several methods, usually based on average sentence length and syllables per word, are included in word-processing programs. For example, Microsoft Word's (Microsoft, 2010) Grammar Checker can calculate the Flesch–Kincaid Grade Level Readability and the Flesch Reading Ease Score. The latter ranges from 100 to 0 (*Very Easy* = 90–100; *Easy* = 80–90; *Fairly Easy* = 70–80; *Standard* = 60–70; *Fairly Difficult* = 50–60; *Difficult* = 30–50; and *Very Difficult* = 0–30), and authors are advised to aim for scores between 60 and 70. Typical psychological reports, even the relatively readable "Summary," attain readability scores within the "very difficult" range (Harvey, 1989, 1997).

The following excerpt is from a report summary quoted in a previous publication (Harvey, 1989). With a Flesch Reading Ease score of 20.1 and a Flesch–Kincaid Grade Level of 15.0, it is far above the recommended level.

> Results of testing present the picture of an angry child whose high need for dominance interferes substantially with his ability to form significant in-depth emotional attachments. Steven dislikes many people and feels justified in aggressing against them. Steven tends to deal with affective issues by blocking them out or ignoring their existence. This denial requires considerable psychological energy, however, which would be better invested in other ways. Steven's emotional status is serious and, without attention, may deteriorate. If Steven seems to be deteriorating behaviorally and the above mentioned suggestions are initiated and maintained for a length of time without success, Steven's educational needs should be re-evaluated with a more restrictive setting considered.[1]

This passage can be rewritten to convey the same information at a much less difficult level. The paragraph below has a Flesch Reading Ease score of 63.2 and a Flesch–Kincaid Grade Level of 8.

> Test results suggest that Steven feels angry and has a high need for control. These feelings cause him to have trouble feeling emotionally close to his family. He also has trouble making friends. Steven dislikes many people and feels that it is all right to be aggressive toward them. He tries to block out or ignore his feelings, which wastes a great deal of his mental energy. Steven's emotional state is serious and may worsen if not addressed. After the above recommendations have been implemented for a few months, his progress should be reviewed. If Steven's behavior has not improved even with these interventions, he may need to be assigned to a special class with fewer students.

Technical terms such as *working memory* and *processing speed* are not familiar to nonpsychologists, and psychologists themselves do not have a common understanding of the definitions of terms such as *learning disability* and *average intelligence* (Harvey, 2006). Although the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; *DSM–IV–TR*; American Psychiatric Association, 2000) has attempted to provide clear definitions for psychiatric conditions, considerable confusion remains over many basic terms, including

[1]From "Eschew Obfuscation: Support Clear Writing," by V. S. Harvey, 1989, *Communiqué, 17*(6), p. 12. Copyright 1989 by the National Association of School Psychologists. Reprinted with permission.

the definition and measurement of intelligence (Sattler, 2008). Furthermore, there is little consistency in defining children as disabled across communities because federal legislation, such as the IDEA (2004), provides considerable latitude to states and local educational agencies in establishing special education eligibility criteria.

There are several possible reasons that psychologists persist in writing reports at a level that consumers cannot easily read. One is that psychologists mistakenly think that writing must be dense in order to gain respect. Another is that psychologists are not taught to write effectively in their graduate programs; even model reports in textbooks used by psychologists in training are written at a "very difficult" readability level (Harvey, 2006), with mean Flesch Grade Level Readability scores above 18.5. Regardless of the reason, it is important that psychologists ensure the readability of their reports if they are to succeed in communicating test results effectively.

To increase the general readability of their writing, psychologists can shorten sentence length, omit passive verbs, and increase the use of subheadings. They can greatly increase the readability of psychological reports by minimizing the use of psychological terms and by providing ecological validity for those terms that are included. This process entails providing concrete examples, preferably from the client's own life (Groth-Marnat, 2009b). For example, a psychologist might define *poor working memory* as "John's difficulty remembering written instructions long enough to complete a multistep task without going back to reread the directions repeatedly." As Ownby (2009) indicated, linking middle-level constructs to data or behavioral descriptions enables the psychologist to mitigate the effect of jargon.

Writing psychological reports at a readable level requires deliberate effort. Psychologists attempting to improve the clarity of their reports are likely to find it necessary to monitor the difficulty level of their reports and edit them to be at a more readable level (Harvey, 1997). Psychologists can easily calculate the reading level of their reports using the Grammar Checker on their word-processing program. They also can solicit consumer feedback (Ownby, 1997). In addition, while editing a report, it is very helpful to imagine oneself as the report's nonpsychologist consumer. Ritzler (1998) suggested revising psychological reports as though writing for grandmothers, "intelligent women who know little about psychology, but who can be sensitive and empathic when they understand someone's personality" (pp. 422–423).

Report Length

Psychological reports can vary enormously in terms of length; Ackerman (2006) defined brief reports as two pages or less and comprehensive reports as 30 to 50 pages. However, report length can be problematic. If they are so short that they report only test scores and provide a statement regarding eligibility for services, they do not include useful information that will encourage the improvement of client functioning. If they are extremely long and all inclusive, they are likely to go unread or readers will have difficulty determining which information is the most significant (Surber, 1995; Tallent, 1993).

The desired length of a psychological report depends on the context in which it is presented. Although medical professionals may desire a one-page bulleted "professional letter" report (Ownby, 2009), the typical length of a psychological report in clinical settings is five to seven pages (Groth-Marnat & Horvath, 2006). However, it is challenging to include all pertinent information and data in a report of this length.

One effective method to shorten reports is to include minimal quantitative data in the body of the reports and, instead, append them in data summary sheets. In addition to shortening the report, this practice has the advantage of removing numerical information that is frequently misunderstood by readers, but it still allows inclusion of quantitative data that succinctly communicate to other psychologists and hold the psychologist accountable (Groth-Marnat & Horvath, 2006). In addition, psychologists might consider streamlining reports using bulleted information and tables of data rather than narratives. This approach has been found to take less time, have no effect on consumer satisfaction, and increase consumer understanding (Dunham, Liljequist, & Martin, 2006).

Tool Selection

Perhaps it is stating the obvious, but judiciously selecting assessment tools and computer software and using them appropriately is indispensible to having meaningful test results to report. Appropriately selecting assessment tools and processes is discussed extensively in other chapters, and this chapter's discussion focuses on tool selection only insofar as it affects reporting test results.

Professional associations have provided detailed ethical and practice guidelines regarding assessments, both individually (American Counseling Association, 1995; American Psychological Association [APA]; 2002; National Association of School Psychologists, 2010) and collaboratively in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999). These professional groups concur that psychologists should not use instruments that have been discredited or insufficiently normed (see Norcross, Koocher, & Garofalo, 2006). Instead, psychologists should exclusively use reliable tests and administer, score, and interpret them as designed to minimize diagnostic errors (Alfonso & Pratt, 1997; Dumont & Willis, 2003). Instruments, and their associated constructs, should be within the training and professional development of the psychologist using them (Harvey & Struzziero, 2008), valid for the client and referral problem, and normed on a genuinely representative, sufficiently large sample. This sample should have been appropriately stratified for gender, geographic regions, racial and ethnic groups, disabilities, income, and education. Further, norm-referenced tests should be used only when the normative data include the population in which the client is a member.

These criteria are particularly challenging to satisfy when working with clients whose native language is not English. Even if an instrument is available in the client's native language and was normed on an appropriate population, psychologists should remember that testing English-language learning clients in only one language provides limited pictures of knowledge, skills, abilities, and instructional needs because their knowledge bases may vary by language. For example, a native Spanish speaker may be Spanish dominant but only understand science in English if all exposure to science had been in English. Although preferable to assessing a client in a language in which he or she is not proficient, nonverbal cognitive scales should be used with caution because nonverbal tests are not entirely culture free, subscales do not represent typical classroom tasks, and even minimal verbal directions are likely to include abstract terms (e.g., prepositions) that are particularly difficult for language learners (Esquivel, Lopez, & Nahari, 2007; Ortiz, 1997). To augment problematic tests, psychologists can explore English-language learners' learning processes. For example, when a dynamic assessment is conducted, three versions of a task, such as Matrix Reasoning, are administered; the first as a pretest, the second to teach the task and observe learning approaches and teaching methods, and the third as a posttest (Lidz, 2003). For a detailed treatment of dynamic assessment, interested readers should consult Volume 3, Chapter 7, this handbook.

Traditional assessments tend to use the same assessment battery regardless of referral question and almost invariably include a standardized cognitive assessment tool. This traditional "shotgun" approach of using a standard battery for every client, including a standardized cognitive instrument as well as an in-depth personality assessment, is not appropriate unless relevant to the referral questions. For example, for many clients, the results of a cognitive scale add little to the diagnostic determination. Such standardized approaches can violate ethical guidelines mandating that client privacy is safeguarded by gathering only the information needed for good decision-making (Sandoval & Irvin, 1990). Administering, interpreting, and writing up unnecessary tests is unjustifiably time consuming for everyone involved and seriously diminishes the resources available for other services.

COMPUTER SCORING AND COMPUTER-GENERATED REPORTS

Completing an entire psychological evaluation takes from 4 to 24 hours with a median of 11.7 hours (Lichtenstein & Fischetti, 1998). Writing a report takes a novice approximately 7 hours and an

experienced psychologist approximately 3 hours (Whitaker, 1994). Writing easily understood, consumer-friendly, contextually based reports takes even more time because of increased consultation with others (Brenner, 2003; Sattler, 2008) and the need to take time to revise them to achieve a more readable level (Harvey, 1997). These time requirements cause considerable frustration for psychologists, who perceive that paperwork seriously reduces the time they have available to provide prevention and intervention services (Harvey & Pearrow, 2010). It also increases the time lag between referral and sharing the results with referring sources, thereby reducing treatment adherence (Meichenbaum & Turk, 1987).

One method to significantly reduce the time it takes to generate a psychological report is to use computer test-scoring programs and computer-generated reports, as so doing can cut the writing time required in half (Ferriter, 1996). However, responsible psychologists use computer report-writing programs with extreme caution for several reasons. Using computer-generated reports is explicitly identified as a highly questionable practice by experts in both testing and ethics. Psychologists may mistakenly attribute computer-generated information with greater accuracy than it deserves (Matarazzo, 1985), may be tempted to use instruments beyond their level of expertise (Carlson & Harvey, 2004), and may assume that computer-generated reports are valid when in fact their validity is dependent on the accuracy of the clinicians who helped formulate the programs (Snyder, 2000).

To mitigate these factors, psychologists should make informed judgments before purchasing test-scoring and report-writing software by consulting published reviews, critically examining the models on which computer programs are based (Maddux & Johnson, 1993; Moreland, 1992), and assessing the credentials of the "virtual consultant" embedded in the software (Carlson & Harvey, 2004; Moreland, 1992). Because ultimate responsibility for administration, scoring, results interpretation, and recommendations reside with the psychologist rather than the computer software, psychologists should also ensure that appropriate use of the software lies within their levels of competence in terms of both the tool and the psychological construct (APA, 2002; NASP, 2010). They should also conceptualize the generated information as originating from a consultant, rather than from an infallible source, and weigh its appropriateness accordingly (Matarazzo, 1985; Ownby, 1997).

Software programs that score tests, interpret the results, generate reports, and suggest diagnoses have the potential to reduce the endemic low interrater reliability that results from diagnoses being based entirely on clinicians' insight, experience, and interpretation (Pardeck, 1997) by essentially providing access to experts whose knowledge has been incorporated into the program's analytical structures (Anastasi & Urbina, 1997; Carlson & Harvey, 2004; Kamphaus, 1993; Moreland, 1992; Snyder, 2000). For example, software included in *Essentials of WISC–IV Assessment* (2nd ed.; Flanagan & Kaufman, 2009) and *Essentials of Cross-Battery Assessment* (2nd ed.; Flanagan, Ortiz, & Alfonso, 2007) helps to determine quickly which scores are validly reported and calculates the Global Ability Index when Full Scale scores are inappropriate because Index scores are too disparate. For clients who are not native English speakers, scoring software included in the Woodcock–Johnson test battery determines the Cognitive Academic Language Proficiency (CALP) or the degree of language proficiency necessary for success in English-based learning situations. These benefits are most evident when computer programs are used to assist in scoring rather than in generating recommendations.

A very serious concern with computer-generated information is that it does not interpret test data within complex contextual variables such as cultural, economic, educational, and social factors (Harvey, Bowser, Carlson, Grossman, & Kruger, 1998; Ownby, 1997). Therefore, when used in isolation, computer-generated reports do not meet the *Standards for Educational and Psychological Testing* (AERA et al., 1999), which mandate that data from multiple sources be aggregated before making significant decisions about individuals. Before finalizing a report that is based on computer software, psychologists should carefully delete hypotheses and recommendations that are not supported by clinical judgment, replace jargon with language

understandable to nonpsychologists, revise irritatingly "canned" narratives, and integrate the results with contextual information (Eyde et al., 1993; Harvey et al., 1998; Harvey & Carlson, 2003; Ownby, 1997). As Litchenberger (2006) indicated, psychologists must carefully sift the overabundance of information generated by computer-based software; not every point produced by computer software is worthy of being included in the final report.

Written Report Components

The current practice of psychology rightfully emphasizes evidence-based practice. In turn, an increased emphasis on evidence-based practice expands traditional psychological report components to include methods previously used to attempt to solve the problem, problem identification and analysis, baseline data, problem and goal definitions, methods to use in intervention implementation, and recommended formative and summative intervention evaluation methods (Brinkman et al., 2007). Considerations for incorporating these factors in psychological reports are discussed relative to each component here.

Identifying Information

Fundamental identifying information includes the client's name, birth date, age, assessment dates, report dates, the psychologist's name and credentials, a statement regarding confidentiality, and the source of informed consent.

Reason for Referral

The stated reasons for referral open the report and undergird the entire assessment process: choice of assessment methods, information integration, and intervention selection. Truncated reasons for referral should be expanded on and clarified using information obtained in interviews with clients and others (Ackerman, 2006) until they are clear, specific, and measurable.

For example, a traditional academic-oriented referral question might be, "Theresa was referred to determine whether she has a specific learning disability in reading." As discussed by Rogers (2010), such an academic-oriented referral question can be rewritten to be both clear and answerable as is appropriate with a problem-solving model. For example, the referral question might be expressed as follows:

> Because Theresa's teacher reports that she is not reading as fluently as expected in third grade, this evaluation was requested to determine Theresa's current reading skills and to analyze the factors contributing to her delay. To support Theresa's reading skills, the assessment will recommend interventions that consider the context of her current reading, her previous responses to instructional strategies that have been provided, her interests, and her cognitive resources.

A traditional behaviorally oriented referral question might be, "Albert was referred because his behavior is disruptive and to determine whether the sheltered workshop work environment is appropriate." Rewritten, the referral questions might be,

> Albert seems to have difficulty engaging in workshop activities. During the day he wanders around the room, refuses to complete assigned work, and interrupts others. He also touches, pulls, or grabs others. This evaluation was requested to compare Albert's behavior to peers, analyze factors that lead to his successes or difficulties, assess the match between his skills and the workshop environment, and suggest interventions to help him behave more appropriately.

Although stated at the beginning of the report, when using a problem-solving approach, the reasons for referral are repeatedly revisited and progressively refined. For example, the initial reason for referral may be to investigate whether a college student is eligible for test accommodations as a student with a learning disability. During initial information gathering (interviews and file reviews), the psychologist discovers that the student does not complete reading assignments and does not have writing skills sufficient to write term papers. This discovery leads to a refinement of the problem definition and problem analysis to include investigations regarding behavior, motivation, social support networks, and

assignment difficulty. During individual sessions with the student, the psychologist observes symptoms of depression, which leads to a further refinement of the problem definition and problem analysis regarding the duration and severity of the depression. Throughout this process, all findings are integrated and contextualized in light of the client's self-reports, performance, educational history, family context, and medical history including vision and hearing acuity. Initial referral questions are expanded on and integrated with concerns raised by others, findings of previous evaluations, and current information gained from multiple sources to shed light on commonalties and disparities.

Background Information

A critical question underlying every psychological assessment is whether the client's current difficulties are situation specific or indicative of persistent or ubiquitous patterns of behavior. Test results provide a snapshot of behavioral responses in one setting and at one point in time. To determine whether test results generalize validly to other settings and to ensure that the reader has a context within which to place test results, the psychologist describes the client's educational, familial, cultural, linguistic, medical, and occupational background using data gathered and aggregated from file reviews, observations, and interviews with the client and others.

In reporting background information, the psychologist takes care to include only accurate information and also consistently cites information sources. Furthermore, according to Bersoff (1995) and St. George and Wulff (1998), clients (and parents, when the client is a child) should be consulted regarding which background information is included in psychological reports and which information is kept private, taking into consideration the intended use of test results and who will have access to the report. This process clearly depends on the purpose of the report; for example, this consultation would be inappropriate in custody determination assessments (see Chapter 34, this volume). However, for most reports, it is highly preferable as it conveys both respect and sensitivity.

Educational background information includes the client's formal education, academic successes and difficulties, grade retention, number of schools attended, remedial or transition programs, special education programming, and school attendance. For clients currently in school, additional information such as patterns of strengths and weaknesses, study habits, and success on high stakes tests also may be important.

Work background information includes the client's work history, challenges, and successes.

Medical background includes any unusual developmental milestone history as well as the health history (past and current serious illnesses, allergies, recurrent ear infections, high fevers, accidents, injuries, physical problems, medications, and medical interventions). Vision and hearing acuity test results should be included as they affect the ability to respond to test stimuli.

Family background includes family composition, family members' health, educational levels, and occupations. When the client is a child, the parent or parent-surrogate's perception of a child's referral problem, additional parental concerns, and parental perceptions of a child's strengths, hobbies, interests, social skills, and autonomy are also important.

Cultural background information may include country of origin, familial culture, extent of the family's or client's acculturation, and the family's use of community supports.

Linguistic background is essential whenever a client's first language is not the dominant language of the setting in which he or she is expected to function. The psychologist should clearly state the linguistic skills of the client and indicate the justification for the language of testing. This justification should take into account the fact that acquiring CALP in a second language takes 5 to 7 years to accomplish and is dependent on many factors, including language of instruction, language spoken at home and in the community, and acculturation. Before identifying a nonnative English speaker as having a language-related learning disorder such as a communication disorder, learning disability, autism spectrum disorder, or intellectual developmental disorders, the psychologist must determine the language proficiency and preference in four activities: listening, speaking, reading, and writing. This is accomplished through assessments by English

as a Second Language specialists, questionnaires, rating scales, observations, and reviews of oral and written language samples both in English and in the native language. A student attending an English-speaking school or college but whose first language was not English should not be identified as having a disability unless that disability is evident in the native language as well as in English (Ortiz, 1997).

Behavioral Observations

In this section, the psychologist describes relevant client behaviors during the assessment sessions such as rapport, anxiety, mistrustfulness, attentiveness, persistence, and self-confidence. It is helpful for the psychologist to give specific and concrete examples both to illustrate the behaviors and to describe behavior variations in response to different activities.

Some psychologists describe a client's physical appearance and manner of dress in order to convey the client's cultural background, economic condition, and care (Nuttall et al., 1999), but others consider the practice archaic (Ownby, 1997). If a physical description is included, it should be critical to the client's issues and as objective as possible. However, even seemingly objective descriptors have acquired idiosyncratic or pejorative meanings and thus should be used with care. For example, one psychologist had the experience of describing a client as "blond," meaning light-haired, but was interpreted as intending to suggest that the client was naive and unintelligent.

In this section, the psychologist also reports testing modifications such as using an interpreter or retesting earlier failed items after a test is completed to see what the client can accomplish with additional support. Examples of such "testing the limits" include suspending time restraints for timed tests, contextualizing vocabulary words by asking the client to use them in a sentence, encouraging the client to use paper and pencil to solve arithmetic problems, and using a test–teach–retest format for items the client might not have been exposed to before testing.

A statement of the probable reliability and validity of the results concludes this section. This statement is critically important. If rapport, cooperation, or perseverance is perceived to be so poor, or if anxiety is so high, that the test results are considered invalid, the psychologist should refrain from reporting the scores in the Results and Interpretation section. If the psychologist does report and include invalid scores, readers are unfortunately likely to focus on them more than the cautionary statements. Thus, they are best left unreported.

Results and Interpretation

In this section, the psychologist describes the outcomes of both standardized and nonstandardized instruments. As suggested earlier in this chapter, it is more helpful to consumers when this section is organized by themes that reflect referral questions. It is also most helpful when it integrates information from multiple sources rather than being confined to tools used by the psychologist.

Results of others' assessments should be reported, attributed, and integrated as appropriate, as so doing can greatly facilitate interpretation of strengths and weaknesses. For example, reporting results on instruments such as the Woodcock–Johnson Achievement Scales, even when administered by another professional, provides the psychologist with opportunities to compare achievement test results with scores on relevant cognitive scales. It also permits the derivation of a rating of a client's CALP to determine English-language proficiency.

Although only current test results are typically included in this section, Ritzler (1998) recommended that previous test results be incorporated as well; doing so facilitates comparison with current results. It also avoids prejudicing the reader's interpretation of current results by their first reading outdated information.

Theme selection. When the themes used to organize a report are meaningful to the reader, they greatly facilitate test interpretation. Therefore, they need to be selected carefully. For example, the Cattell–Horn–Carroll theory can be used in eligibility and program decision-making (Fiorello & Primerano, 2005) and is operationalized in some standardized batteries, such as the Woodcock–Johnson Battery (McGrew & Woodcock, 2001). In addition, the widely used Wechsler scales (Wechsler, 2002, 2003, & 2008) can be interpreted in light of

this theory (Flanagan & Kaufman, 2009; Flanagan, Ortiz, & Alfonso, 2007). However, as described by Gomez (2006), using this theory's domains (fluid reasoning, quantitative ability, visual processing, crystallized intelligence, processing speed, short-term memory, long-term retrieval, and auditory processing) as the foundation for theme-based reports does not improve consumer understanding.

Instead, themes that reflect referral questions are most helpful (Groth-Marnat, 2009a). In the previous example of a referral question regarding "Theresa," a theme-based report might be organized to answer the four questions: (a) What are Theresa's reading skills at the word level (i.e., phonemic awareness, automatic naming of individual words, word decoding, and cognitive fluency)? (b) What are Theresa's skills in reading comprehension and its associated learning abilities (verbal reasoning, language, listening comprehension, working memory, and perceptual reasoning)? (c) How persistent is Theresa as she approaches reading or learns other material? (d) How has Theresa responded to reading instruction, both historically and in her current classroom? (Rogers, 2010).

Similarly, in response to the previous example of a referral questions regarding "Albert," a theme-based report might be organized to answer the following: (a) How do Albert's basic skills compare with the skill levels required for success, and does he have areas in need of extra support? (b) Do Albert's behaviors reflect difficulties in language, verbal reasoning, cognitive flexibility, memory, or fluid reasoning? (c) How does Albert's behavior compare with that of peers, and what environmental factors provoke or sustain his problem behaviors? (d) What interventions have been tried, and how has Albert responded? (Rogers, 2010). Relevant data to answer these questions might include current and former performance evaluations, interventions attempted, and information from progress monitoring.

Results from client interviews, incorporated into relevant themes, also enrich findings. These can convey the client's perceptions of the problem; special abilities, talents, interests, favored activities, friends and preferred work partners; what he or she does that makes others feel happy, sad, or angry; significant persons, including losses; adjustment to a new culture; leisure activities; and future dreams and vocational aspirations.

For clients with behavioral concerns, data regarding functional behavioral analyses and assessments are also essential. These report the results of structured behavior observations in natural settings. Critical components include a definition of the problem behavior in specific, measurable, and easily understood terms; data regarding the behavior from multiple observations; observed antecedents, and consequences that seem to maintain inappropriate behavior; and data regarding hypothesis elimination.

Score reporting. Traditional psychological reports incorporate, within the body of the report, a list of the assessment procedures used, descriptions of each test instrument, and scores obtained. However, in the interest of making reports more accessible to nonpsychologist readers, and because scores are so commonly misunderstood, psychologists can place these elements, including data tables, at the end of the report as appendixes, as suggested earlier in this chapter.

Regardless of placement, the scores should be reported with care because test scores are meaningless unless contextualized. As stated in the *Standards for Educational and Psychological Testing* (AERA et al., 1999),

> Test scores, per se, are not readily interpreted without other information, such as norms or standards, indications of measurement error, and descriptions of test content. Just as a temperature of 50° in January is warm for Minnesota and cool for Florida, a test score of 50 is not meaningful without some context. (p. 62)

In both oral and written reporting, the psychologist clearly explains standard scores, scaled scores, normal curve equivalents, percentile ranks, and other scores. For example, scores presented as standard scores should be accompanied by percentile scores and an explanation thereof ("Suzy's standard score of 100, which falls at the 50th percentile, indicates that out of a group of 100 persons, she would usually score higher than about 49 and lower than about 49").

In respect for reliability and the standard error of measurement, scores should be reported using 95% confidence bands. Psychologists should clearly explain the meaning of these confidence bands in terms that a nonpsychologist can understand and indicate that the confidence band is inherent in the test process and does not take into account errors or test administration problems (Dumont & Willis, 2003). It is also helpful, as a rule, to explicitly encourage caution regarding any obtained scores because they are indicative of functioning in a specific time and place. Finally, in respect for measurement error, psychologists generally refrain from using rigid cutoff scores for decision-making.

In respect for validity, psychologists should interpret tests only insofar as validity evidence exists to support specific uses and interpretations of test scores. They should also clarify that subtests, factors, indices, and tests with similar or identical titles can vary greatly. It is also helpful when psychologists make clear statements linking findings to implications so that readers can associate results with suggested interventions and programs.

Summary and Diagnostic Impressions

The summary clarifies the answers to the original referral questions, integrates the findings of the entire process, and draws conclusions on the basis of multiple sources. The reader of the full report should not be surprised by any element in the summary. Its content should logically follow the rest of the report.

Many assessments are conducted to determine whether a client is eligible to receive services as a result of a disability under IDEA (2004) or a diagnosis according to the *DSM–IV–TR* (American Psychiatric Association, 2000). However, psychologists should take great care in concluding that assessment results indicate a disability diagnosis because such diagnoses have such profound implications for clients' futures. Although a disability designation enables a client to access mental health, special education, and other services, it can also severely restrict clients' future options and thereby infringes on their autonomy, which in turn can violate the ethical principle of beneficence and autonomy (Michaels, 2006). Furthermore, meta-analyses have indicated that some placements have negative rather than positive effects (Kavale, Forness, & Siperstein, 1999; Sheridan & Gutkin, 2000) and that minority group members are placed in such placements disproportionately (Reschly, 2006). Recommending programs that are not beneficial, much less those that are harmful, clearly violates the ethical mandates to "do no harm" and provide "equal protection."

Psychologists should also be aware that, when working with children, the criteria specified in the *DSM–IV–TR* may not match the criteria specified in the IDEA (2004) or the state guidelines for IDEA implementation. For example, the IDEA definition of *severe emotional disturbance* excludes students who are socially maladjusted unless they also meet criteria for emotional disturbance. This means that children identified as having oppositional defiant disorder and academic deficiencies may not meet criteria for special education services unless another disorder such as depression or anxiety is present.

Furthermore, current perceptions of disabilities are that they are developmental and contextual rather than fixed and intrinsic to the individual. As indicated by the American Association on Intellectual and Developmental Disabilities (2002), when individuals with intellectual developmental disorders receive appropriate environmental supports for their limitations, life functioning improves over time. Similarly, contemporary identification of children with learning disabilities contextualizes this identification relative to their responses to instruction and intervention (Mather & Gregg, 2006).

Recommendations

Psychological reports can be rendered ineffective when they include vague recommendations that would be applicable to any client, that are inappropriate because they do not match the assessment results or the client's context, or because they are not presented in a manner that facilitates treatment adherence or integrity of intervention implementation. Furthermore, the high investment of resources involved in an evaluation calls for more than the most obvious recommendations. Recommendations should be specific, clear, and evidence based, and each recommendation should be linked to a reason for referral and specific assessment findings.

Consumers must perceive that recommended interventions are appropriate, effective, and possible to implement. Persons responsible for implementing each intervention should be identified to increase accountability; recommendations without identified implementers invites their being ignored because of the tendency to assume that "someone else" will implement and monitor them. Interventions that are highly complex or require substantive lifestyle changes are less likely to be put in place (Meichenbaum & Turk, 1987). Whenever possible, recommended interventions should be built on existing structures and include a plan for progress monitoring on a regular basis, as monitoring leads naturally to intervention modification and maximizes client progress. Optimally, plans for moving to generalization to other settings as well as client self-monitoring are included. To meet these characteristics, psychologists must be well enough informed about the clients' environment to know that the recommended interventions are appropriate.

It is also helpful to give consumers choices in terms of intervention selection and structure and to convince them that implementing the interventions will have benefits that outweigh their inconvenience. To these ends, it is very helpful to include the client and others (e.g., life partners, family members, teachers, and parents) as collaborators so that they can provide information regarding the appropriateness of various interventions; tentative recommendations can be brought to meetings but not formalized until collaboration with other interested parties has occurred. The most effective recommendations are accompanied by handouts, training, monitoring implementation integrity, and progress monitoring.

Finally, recommendations should consider clients' strengths as well as weaknesses, deficiencies, or disabilities. Snyder, Ritschel, Rand, and Berg (2006) advocate using a robust predictor of psychological health, Hope Theory, to balance the typically negative perspective in psychological reports. After positive, as well as problematic, information is gathered from the client and others to obtain a complete view of the client, strengths can be used to help develop goals, choose pathways, and empower the client's agency. Such incorporation of athletic, mechanical, musical, social, creative, and other interests, skills, and talents into recommendations can move the assessment beyond a pathological focus and lead to strength-based interventions.

CONCLUSION

Psychologists should strive to communicate test results so clearly and persuasively that recipients of the information are motivated to use suggested changes. Doing so requires considering contextual variables, selecting tools carefully, organizing the report carefully, writing in a readable fashion, and modifying the report components. As Groth-Marnat (2009a) stated,

> Based on research, practice, and teaching . . . five crucial features would greatly improve psychological reports: increase readability, connect interpretations to the person's context, integrate interpretations around relevant domains, include client strengths, and provide clear links between the referral questions. (p. 303)

References

Ackerman, M. J. (2006). Forensic report writing. *Journal of Clinical Psychology, 62*, 59–72. doi:10.1002/jclp.20200

Alfonso, V. C., & Pratt, S. I. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 326–344). New York, NY: Guilford Press.

American Association on Intellectual and Developmental Disabilities. (2002). *AAIDD definition of mental retardation.* Retrieved from http://www.aaidd.org/content_104.cfm

American Counseling Association. (1995). *Code of ethics and standards of practice.* Alexandria, VA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073. doi:10.1037/0003-066X.57.12.1060

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Bersoff, D. N. (Ed.). (1995). *Ethical conflicts in psychology*. Washington, DC: American Psychological Association. doi:10.1037/10171-000

Blais, M. A., & Smith, S. R. (2008). Improving the integrative process in psychological assessment: Data organization and report writing. In R. Archer & S. Smith (Eds.), *Personality assessment* (pp. 405–439). New York, NY: Taylor & Francis Group.

Brenner, E. (2003). Consumer-focused psychological assessment. *Professional Psychology: Research and Practice, 34*, 240–247. doi:10.1037/0735-7028.34.3.240

Brinkman, T. A., Segool, N. K., Pham, A. V., & Carlson, J. S. (2007). Writing comprehensive behavioral consultation reports: Critical elements. *International Journal of Behavioral Consultation and Therapy, 3*, 372–383.

Carlson, J. F., & Harvey, V. S. (2004). Using computer-related technology for assessment activities: Ethical and professional practice issues for school psychologists. *Computers in Human Behavior, 20*, 645–659. doi:10.1016/j.chb.2003.10.010

Cuadra, C. A., & Albaugh, W. P. (1956). Sources of ambiguity in psychological reports. *Journal of Clinical Psychology, 12,* 267–272. doi:10.1002/1097–4679(195604)12:2<109::AID-JCLP2270120203>3.0. CO;2-Y

Dumont, R., & Willis, J. O. (2003). Issues regarding the supervision of assessment. *Clinical Supervisor*, *22,* 159–176. doi:10.1300/J001v22n01_11

Dunham, M., Liljequist, L., & Martin, J. (2006). Streamlining psychological reports. *Trainers' Forum: Periodical of the Trainers of School Psychologists*, *25*(4), 9–14.

Esquivel, G., Lopez, E. C., & Nahari, S. G. (Eds.). (2007). *Handbook of multicultural school psychology: An interdisciplinary perspective*. Mahwah, NJ: Erlbaum.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., . . . Primoff, E. S. (1993). *Responsible test use: Case studies for assessing human behavior*. Washington, DC: American Psychological Association. doi:10.1037/10132-000

Family Education Rights and Privacy Act of 1974 (FERPA), Pub. Law 93–380, 20 U.S. C. A. § 1232g, 34 C. F. R. § Part 99 (1993).

Ferriter, M. (1996). Automated report writing. *Computers in Human Services, 12*, 221–228. doi:10.1300/J407v12n03_03

Fiorello, C., & Primerano, D. (2005). Research into practice: Cattell-Horn-Carroll cognitive assessment in practice: Eligibility and program development issues. *Psychology in the Schools*, *42,* 525–536. doi:10.1002/pits.20089

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC–IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

Gomez, H. D. (2006). Teachers' preferences for and comprehension of psychological reports: Test-oriented versus Cattell–Horn–Carroll models. *Dissertation Abstracts International: Section A. Humanities and Social Sciences, 67*(4), 1221.

Groth-Marnat, G. (2009a). The five assessment issues you meet when you go to heaven. *Journal of Personality Assessment, 91*, 303–310. doi:10.1080/00223890902935662

Groth-Marnat, G. (2009b). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Groth-Marnat, G., & Horvath, L. S. (2006). The psychological report: A review of current controversies. *Journal of Clinical Psychology, 62*, 73–81. doi:10.1002/jclp.20201

Harvey, V. S. (1989). Eschew obfuscation: Support clear writing. *Communiqué*, *17*(6), 12.

Harvey, V. S. (1997). Improving readability of psychological reports. *Professional Psychology: Research and Practice, 28*, 271–274. doi:10.1037/0735-7028.28.3.271

Harvey, V. S. (2006). Variables affecting the clarity of psychological reports. *Journal of Clinical Psychology, 62*, 5–18. doi:10.1002/jclp.20196

Harvey, V. S., Bowser, P., Carlson, J. F., Grossman, F., & Kruger, L. (1998, April). *School psychologists and high technology: Ethical dilemmas and considerations.* Symposium conducted at the National Association of School Psychologists Convention, Orlando, FL.

Harvey, V. S., & Carlson, J. (2003). Ethical and professional issues with computer related technology. *School Psychology Review, 32*, 92–107.

Harvey, V. S., & Pearrow, M. (2010). Identifying challenges in supervising school psychologists. *Psychology in the Schools, 47*, 567–581. doi:10.1002/pits.20491

Harvey, V. S., & Struzziero, J. A. (2008). *Professional development and supervision of school psychologists: From intern to expert.* Thousand Oaks, CA: NASP and Corwin/Sage.

Individuals With Disabilities Education Improvement Act of 2004, Pub. L. No. 108–446, 118 Stat. 2647 (2004).

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston, MA: Allyn & Bacon.

Kavale, K. A., Forness, S. R., & Siperstein, G. N. (1999). *Efficacy of special education and related services.*. Washington, DC: American Association on Mental Retardation.

Koocher, G. P., & Keith-Spiegel, P. (1998). *Ethics in psychology: Professional standards and cases* (2nd ed.). New York, NY: Oxford University Press.

Lichtenstein, R., & Fischetti, B. A. (1998). How long does a psychoeducational evaluation take? An urban Connecticut study. *Professional Psychology: Research and Practice*, *29,* 144–148. doi:10.1037/0735–7028.29.2.144

Lidz, C. (2003). *Early childhood assessment.* Hoboken, NJ: Wiley.

Litchenberger, E. O. (2006). Computer utilization and clinical judgment in psychological assessment reports. *Journal of Clinical Psychology, 62*, 19–32. doi:10.1002/jclp.20197

Maddux, C. D., & Johnson, L. (1993). Best practices in computer-assisted assessment. In H. B. Vance (Ed.), *Best practices in assessment for school and clinical settings* (pp. 177–200). Brandon, VT: Clinical Psychology.

Martin, W. T. (1972). *Writing psychological reports.* Springfield, IL: Charles C Thomas.

Matarazzo, J. D. (1985). Clinical psychology test interpretations by computer: Hardware outpaces software. *Computers in Human Behavior*, *1,* 235–253. doi:10.1016/0747–5632(85)90015–9

Mather, N., & Gregg, N. (2006). Specific learning disabilities: Clarifying, not eliminating, a construct. *Professional Psychology: Research and Practice*, *37,* 99–106. doi:10.1037/0735–7028.37.1.99

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson battery* (3rd ed.). Itasca, IL: Riverside.

Meichenbaum, D., & Turk, D. C. (1987). *Facilitating treatment adherence: A practitioner's guidebook.* New York, NY: Plenum Press. doi:10.1007/978-1-4684-5359-1

Michaels, M. H. (2006). Ethical consideration in writing psychological assessment and reports. *Journal of Clinical Psychology, 62*, 47–58. doi:10.1002/jclp.20199

Microsoft. (2010). *Microsoft Word 2010 help text*. Seattle, WA: Author.

Moreland, K. L. (1992). Computer-assisted psychological assessment. In M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view* (pp. 343–376). Palo Alto, CA: Consulting Psychologists Press.

National Association of School Psychologists. (2010). *Model for comprehensive and integrated school psychological services*. Bethesda, MD: Author. Retrieved from http://www.nasponline.org/standards/2010standards/2_PracticeModel.pdf

Norcross, J. C., Koocher, G. P., & Garofalo, A. (2006). Discredited psychological treatments and tests: A Delphi poll. *Professional Psychology: Research and Practice, 37*, 515–522. doi:10.1037/0735-7028.37.5.515

Nuttall, E. V., Devaney, J. L., Malatesta, N. A., & Hampel, A. (1999). Writing assessment results. In E. V. Nuttall, I. Romero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers: Psychological and educational dimensions* (pp. 396–406). Boston, MA: Allyn & Bacon.

Ortiz, A. A. (1997). Learning disabilities occurring concomitantly with linguistic differences. *Journal of Learning Disabilities, 30*, 321–332. doi:10.1177/002221949703000307

Ownby, R. (2009). Writing clinical reports. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 684–692). New York, NY: Oxford University Press.

Ownby, R. L. (1997). *Psychological reports: A guide to report writing in professional psychology* (3rd ed.). New York, NY: Wiley.

Pardeck, J. T. (1997). Computer technology in clinical practice: A critical analysis. *Social Work and Social Sciences Review, 7*, 101–111.

Pelco, L. E., Ward, S. B., Coleman, L., & Young, J. (2009). Teacher rating of three psychological report styles. *Training and Education in Professional Psychology, 3*, 19–27. doi:10.1037/1931-3918.3.1.19

Pryzwansky, W. B., & Hanania, J. S. (1986). Applying problem solving approaches to school psychological reports. *Journal of School Psychology, 24,* 133–141. doi:10.1016/0022-4405(86)90005-1

Reschly, D. J. (2006, October). *Response to intervention in general, remedial, and special education.* Paper presented at the fall convention of the New Hampshire Association of School Psychologists, Concord, NH.

Ritzler, B. A. (1998). Teaching and learning issues in an advanced course in personality assessment. In L. Handler & M. J. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 431–452). Mahwah, NJ: Erlbaum.

Rogers, L. (2010, June). *The role of referral questions in a problem-solving approach to assessment.* Presentation at the Massachusetts School Psychology Trainer's Third Annual Supervision Institute, Boston, MA.

Rucker, C. M. (1967). Technical language in the school psychologist's report. *Psychology in the Schools, 4*, 146–150. doi:10.1002/1520-6807(196704)4:2<146::AID-PITS2310040210>3.0.CO;2-9

Sandoval, J., & Irvin, M. G. (1990). Legal and ethical issues in the assessment of children. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 58–78). New York, NY: Guilford Press.

Sattler, J. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

Sheridan, S., & Gutkin, T. (2000). The ecology of school psychology: Examining and changing our paradigm

for the 21st century. *School Psychology Review, 29*, 485–502.

Shively, J. J., & Smith, A. E. (1969). Understanding the psychological report. *Psychology in the Schools, 6*, 272–273. doi:10.1002/1520-6807(196907)6:3<272::AID-PITS2310060309>3.0.CO;2-U

Snyder, C. R., Ritschel, L. A., Rand, K. L., & Berg, C. J. (2006). Balancing psychological assessments: Including strengths and hopes in client reports. *Journal of Clinical Psychology, 62*, 33–46. doi:10.1002/jclp.20198

Snyder, D. K. (2000). Computer-assisted judgment: Defining strengths and liabilities. *Psychological Assessment, 12*, 52–60. doi:10.1037/1040-3590.12.1.52

St. George, S., & Wulff, D. (1998). Integrating the client's voice within case reports. *Journal of Systemic Therapies*, *17*(4), 3–13.

Surber, J. M. (1995). Best practices in problem-solving approach to psychological report writing. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. 3, pp. 161–169). Washington, DC: National Association of School Psychologists.

Tallent, N. (1993). *Psychological report writing* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Wechsler, D. (2002). *Wechsler preschool and primary scale of intelligence* (3rd ed.). San Antonio, TX: Pearson.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Pearson.

Whitaker, D. (1994). *How school psychology trainees learn to communicate through the school psychological report*. Unpublished doctoral dissertation, University of Washington, Seattle.

CHAPTER 4

THE CLINICAL VERSUS MECHANICAL PREDICTION CONTROVERSY

William M. Grove and Scott I. Vrieze

The literature on clinical versus mechanical prediction spans most of the 20th century until the present day. Legion studies have been procured to address the central question of the debate: Once data have been gathered, is it better to allow an expert judge (clinician) to use the data to arrive at a prediction, or is it better to use an algorithmic formula? Other similar questions have been asked (Holt, 1958, 1970), but this chapter is most concerned with the central question. We use as our starting point the 1954 classic by Paul Meehl. Many may be tempted to stop reading at this point, seeing our starting point as clear evidence of our bias on the subject. We can only hope that the reader will continue, as Meehl arguably gave one of the most fair and balanced treatments of the subject to date. His thoughtful analysis, so often cited and so little read, is an exposition of conflict. In fact, most of the book is a defense of those things the clinician can do that the formula cannot, such as formulate hypotheses and causal theories about that which is to be predicted. In the end, however, it seems one important question was at least partially answered. That is, the mechanical formulas, at least by studies conducted at that time, more accurately predicted future events than expert judges did.

Throughout this chapter, we attempt to explain this result and clarify aspects of it. It is not an all-pervasive result dictating that mechanical prediction must be used at all times, for all purposes, or worse, that it is the only ethical way to make predictions. If that is true, then most clinicians are acting unethically, because 98% of clinicians use clinical judgment to arrive at predictions about their clients (Vrieze & Grove, 2009). It has benefits, to be sure, with one being increased expected predictive accuracy. However, before we discuss experimental comparisons of clinical and mechanical prediction, there are some preliminary issues to discuss. We will not obtain valid conclusions from the literature in this area unless we acknowledge the following:

1. Clinical prediction (diagnostic, prognostic, or other framework) has outcomes that are commensurable with, and hence comparable with, output of mechanical predictions.
2. It is clarified what activity of the clinician (e.g., data gathering, data combination) is being compared to the output of mechanical prediction, prognostication, or diagnosis.
3. There are limitations of the literature that impair or prevent a full analysis of the issue.

As it turns out, the literature lends itself best to answering the following question: How accurate, on average, is clinical prediction when compared with mechanical prediction? Many would rather discuss how well clinicians fare (in comparison with mechanical prediction) in conducting activities other than prediction, such as hypothesis generation or data gathering. Alas, this conflates variables that may not be associated with data combination, which, in our opinion, lies at the center of the clinical–mechanical discussion.

William M. Grove has benefited greatly from numerous conversations and memorandum exchanges on this topic with the late and greatly lamented P. E. Meehl.

DOI: 10.1037/14048-004
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

The principal ethical underpinning of this controversy is the obligation of beneficence to the client. It is asserted that the most accurate prediction is best for the client, absent evidence to the contrary. This assumption is necessary to simplify the discussion. Published studies almost never include the utility information needed to conduct a proper decision theoretic analysis anyway (Von Neumann & Morgenstern, 1944). Pretending that the cost of a false positive (predicting positive when the outcome is negative) and a false negative (predicting negative when the outcome is positive) are equal, accuracy of a prediction method becomes equivalent to its utility, and utility analysis is implicit. At any rate, we can see no reason why the ratio of utilities for clinical predictions should be much different from that for mechanical predictions. Hence, this assumption should not bias the clinical–mechanical prediction comparison for one method and against the other.

TERMINOLOGY

Mechanical prediction is a term that encompasses many different types of prediction models. This includes statistical formulas such as multiple regression, using both linear and nonlinear models, logistic regression, or linear or quadratic discriminant analysis. It also includes computationally intensive machine learning algorithms (e.g., random forests or neural nets; Hastie, Tibshirani, & Friedman, 2009). Perhaps even more well known are actuarial tables—contingency tables selected from the total data set, applying division on a chosen predictor and then retabulating the hit rates across the new table. Insurance companies once relied on such tables to predict, for example, policyholder demise.

Sources of information commonly used by both clinicians and mechanical prediction formulas are clinical impressions obtained from unstructured interviews and/or behavior observations, projective and objective personality and/or psychopathology tests, occupational-related testing, and cognitive tests. A seminal contribution made by Meehl (1954) was to confine attention to the method by which the data is combined. That is, the focus is not on gathering data or on what data are being used to make predictions. It is whether the clinician or the formula is better at using the available data to arrive at predictions, whatever those data may be. This can prove problematic for comparison studies, however, as the clinician routinely has more information available. For example, it is impossible to quantify all information available during interview (eye contact, posture, voice pitch, etc.), nor would it necessarily be helpful to the mechanical prediction algorithm. Note that this could very well be a bias in favor of clinical prediction, as the clinician typically has a wealth of information (nonverbal, intuition) that is not readily available to a mechanical algorithm. Such a bias might be small, or it may be considerable. It may also be a bias in favor of mechanical prediction, as clinicians overwhelmed with information may weight poor information highly, and thus vitiate valid conclusions made on valid predictor variables.

SINGLE-CASE PROBABILITY

A *reference class* is the class of individuals (more generally, cases or possible events) to which the present individual belongs and from which a probability can be generated for her or him. Very simply, to estimate the probability that a patient is diseased, one takes the number of diseased individuals in the reference class and divides by the total number of individuals in the class. A straightforward example is the probability that a randomly selected U.S. citizen has schizophrenia. To estimate this, simply take the number of U.S. citizens with schizophrenia and divide by the population of the United States. Of course, no one actually does this. Instead, random samples are ascertained and the true population rate of schizophrenia is statistically estimated from those samples.

Single-case probability is the philosophical problem of defining a probability on the basis of a single event. Intuitively, probability is often explained as a stochastic property of many events or long sequences of events (e.g., flipping the same coin many times to ascertain probability of heads). Imagine a single event. A coin is produced, flipped, and immediately destroyed. What is the probability that it lands heads? How would one estimate this? Is the notion of probability even applicable to this question? In the clinic, it is the issue of having a unique

individual patient before us and estimating their response to treatment. We may have seen patients similar to this one, but we have not seen this one at this time in their life, nor will we ever again. This point is germane, as some defenders of the routine use of clinical prediction have argued the following:

1. The individual patient is unique and cannot be reduced to membership in simple, strict categories, such as other patients like this one.
2. As prominent philosophers of probability (e.g., Reichenbach, 1949; von Mises, 1957), argued in their frequentist philosophies of probability, the probability concept applies to series of observations, not a single observation.
3. The clinician practically always has extra information that is not included in the mechanical prediction algorithm, and this information cannot simply be used to amend the mechanical prediction, because its precise role in prediction is too ill understood to set up a new mechanical prediction system, based on the old plus the new predictors.

If these statements are true, they invalidate the entire procedure of mechanical prediction, as it is based on using information from other, similar, individuals to arrive at a conclusion about the present unique individual. For example, a single patient may belong to many reference classes, such as the following:

1. All present human beings.
2. All males (alternatively, all females).
3. All males (or females) of the same age as the present patient.
4. Class 3, restricted to all clients of mental health case workers.
5. Class 4, restricted to those with a history of similar episodes of the patient's current ailment.
6. Class 5, restricted to all clients with similar histories and evidence of thought disorder + first-rank symptoms of schizophrenia.
7. Class 4, restricted to those with 6 to 8 "V" Minnesota Multiphasic Personality Inventory (MMPI) profile code types (Dahlstrom, Welsch, & Dahlstrom, 1972). Note that this is not a restriction of preceding classes in the same way that Classes 2 to 4 are such.

Note that new classes are often formed by splitting existing classes, but the splitting can take place on the same class in two or more different ways. This means that if Class 5 yields one predicted probability and Class 3 produces another the predictions are potentially overlapping but can serve as undercutting defeaters (Pollock, 1990). Hence choice of Class 5 as reference class can contradict, and be contradicted by, Class 3. In many circumstances one has no knowledge that either Class 3 or Class 5 is "the" most accurate reference class for generating a prediction.

A rigorous treatment of the reference class and single-case probabilities is far beyond the scope of this chapter. The interested reader is referred to Hájek (2007), Reichenbach (1949), Pollock (1990), and Kyburg (1978), although Kyburg has since modified his views.

Intuitively, when trying to estimate an actuarial probability from predictor information, one chooses the reference class that is narrowest (most specific to the patient) while not being so small in number that the probability estimate is unstable (Reichenbach, 1949). These two goals generally conflict, but deriving objective rules applicable to every prediction scenario is well-nigh impossible. In fact, choosing the reference class is at the heart of the prediction problem, and it remains incompletely solved. Present solutions are rife with complex modal logic. For example, Pollock's (1990) *nomic probabilities* are calculated directly from classes of aggregated individuals' situations across counterfactual possible worlds. Despite the erudite nature of solutions from philosophy of probability, there have been practical advances. For example, Pollock (2007) has derived a function that enables one to add a new predictor variable to an existing mechanical prediction algorithm (e.g., add a new predictor to a regression equation). This *Y*-function does not require new validation samples for the full set of predictors (old + new), and represents a rational solution to Meehl's (1954) "broken-leg case."

The broken-leg case is laid out as follows. Professor X has been observed over many years, and it has been determined that his probability of cinema attendance on Saturday nights is 0.99. If asked to predict whether he will attend a movie this evening,

we would rationally and rather confidently predict in the affirmative. However, we have just learned that the professor has broken his leg and is in a hip cast. In light of this new information, it is reasoned that he will go to the movies in his hip cast with a probability of 0. This scenario illustrates the legitimate ignoring of previous probability information (i.e., a mechanical prediction based on a reference class, namely, Professor X's history of moviegoing) in the face of a new fact with (approximately) a known value of 0 or 1.

Suppose the hypothetical actuarial study did not examine the "hip cast" factor in establishing probabilities of Professor X's movie attendance, because he had never before broken his leg. Nevertheless, Pollock's (2007) *Y*-function allows a conclusion based on probability theory that the professor will, with high probability, skip the movies next Saturday night. The broken-leg case is meant to be a simple illustration, and many would agree that changing a mechanical prediction based on Professor X's history is obvious now that he has broken his leg. However, there are myriad other predictive situations where new information has come to light and was not included in the mechanical prediction algorithm. Some of these scenarios will be far from intuitive, and Pollack's *Y*-function allows a rational way to adjust mechanical prediction outputs in the face of changing circumstances. Appeal to clinical intuition and common sense is not necessary. Pollock's (2007) result is extraordinary, and we refer the interested reader there for more details.

To end this section, consider two thought experiments intended to suggest that single-case probabilities are useful constructs, despite philosophical quandaries. First, suppose there are two identical, never-fired .38 caliber revolvers, identical in all respects except their loading. Each revolver has six chambers for ammunition. One revolver is loaded with a single live round. The other is loaded with five live rounds. You are forced to play one round of Russian roulette with one of these guns after spinning the cylinder. After one trigger pull with the chosen revolver held to your temple, both revolvers will be destroyed. Thus, the thought experiment is constructed so that the likelihood of dying is a single-case probability. The revolver will only be shot once (like our coin only flipped once and then destroyed). Which revolver do you want to use? If you choose the revolver with a single live round, it seems you may be committed to some notion of single-case probability, as you see it as less likely (or some such term) to kill you.

The second thought experiment is more complex, but the exposition simpler. Do you buy insurance? Your insurance premium is based on comparing yourself (a unique individual of which there is, has been, and will be, no other) with others like you. The insurance company essentially makes a bet about you, such as your life expectancy or odds of being in an accident. Insurance companies are profitable, yet they predicate their entire existence on single-case probabilities. If single-case probabilities are nonsensical absurdities, how do insurance companies reliably profit? The answer is that among other skills they may have, these companies make accurate predictions.

THE LITERATURE ON COMPARATIVE ACCURACY OF CLINICAL VERSUS MECHANICAL PREDICTION

The literature on this subject is extensive although, relative to some domains, manageable. We break the review into two sections: single studies that contain unique study designs or were particularly influential in the literature and meta-analytic review studies. The former are thought-provoking seminal articles. The latter are the best evidence so far about the comparative accuracy of these prediction methods.

Table 4.1 contains a list of all studies contained in this review, along with a short description of the study and findings. The table is not meant to offer a comprehensive list of all studies today—far from it. Instead, we highlight some of the more interesting, influential, and comprehensive studies to date. Interested readers are referred to the more recent meta-analyses for comprehensive lists of studies (e.g., Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000).

Single Studies

We begin with three selected articles. Sarbin (1943) investigated the prediction of college grade point

TABLE 4.1

Evaluation of Study and Review Outcomes

Study	Outcome
Sarbin (1943)	Single study. Used high school rank and college board exam scores to predict college grade point average for 162 college freshmen (not cross-validated).
Goldberg (1965)	Single study; *N*s = 961 and 29 judges; mechanical prediction (not cross-validated) superior to average judge and even better than most accurate judge.
Meehl (1954)	Box score = 20 to 0, in favor of mechanical prediction over statistical prediction; some studies give clinicians variables not in mechanical equation.
Holt (1958)	Narrative review of methodology of selected clinical and clinical-versus-statistical studies.
Goldberg (1968)	Correction of Lindzey's statistical error, leading to change of Meehl's box score from 20–1 to 21–0.
Sawyer (1966)	Narrative review and box score of 45 clinical vs. mechanical comparison studies.
Holt (1978)	Amplified narrative review (from 1968) with critiques of proactuarial studies.
Korman (1968)	Review of clinical vs. mechanical prediction of managerial performance; clinician superior to mechanical prediction in some instances.
Sines (1970)	Highly negative review of Sawyer (1966).
Grove and Meehl (1996)	Box scored brief narrative reviews from preliminary data set of Grove et al (2000).
Holt (1986)	New theory of clinical judgment as stemming from statements in the narrative form.
Dawes et al. (1989)	Summary of literature with sampled studies, using Grove et al. (2000) preliminary results as empirical basis, transformed from effect sizes to a box score.
Sarbin (1986)	Narrative review with proclinician arguments.
Grove et al. (2000)	Meta-analysis of 137 comparison studies. Mechanical prediction led to 10% improved accuracy over clinical prediction.
Ægisdóttir et al. (2006)	Meta-analysis of 92 effect sizes from 67 studies confined to mental health predictions. Mechanical prediction led to 12% improvement over clinical prediction. In most stringent subsample of studies (48 effect sizes), mechanical prediction led to 13% improved accuracy.

average (GPA) by high school guidance counselors, given a severely constrained range of predictors. In the era when Sarbin's study was conducted, students were conceptualized as either "college material" or not. Five clinicians were given the high school ranks and college board scores of 162 students, and then the counselors made predictions with no more information available about students or counselors. Finally, Sarbin compared the clinical predictions to a two-variable linear regression equation. No significant difference between these two data combination methods' accuracies was found. Gender was a moderator variable, with women more predictable. For men, the clinical and statistical predictions had multiple R^2 of .35 and .45, respectively. For women, the corresponding coefficients were .69 and .70, respectively.

Holt (1958) severely criticized Sarbin (1943) on the grounds of poor ecological validity for information that clinical psychologists usually possess and from which they make predictions, and as a poor match to typical clinical criteria. Holt ignored the compensating fact that, with clinicians only knowing all the predictors used in mechanical prediction, confounds are balanced out, leading to a less biased comparison of the two classes of prediction procedures.

In contrast to modern diagnoses based on the *Diagnostic and Statistical Manual of Mental Behaviors* (American Psychiatric Association, 2000), all anxiety disorders, seasonal depression, dysthymia, and some personality pathology were all classed together under the rubric "neurosis" by Goldberg (1965). His 861 subjects were diagnosed by the 29 clinicians (14 clinical psychologists and 15 gradual students) into the classes "neurotic" or "psychotic" by degree from MMPI raw or T scores, excluding Mf and Si. It was found that the best statistical model using MMPI validity and clinical scales as predictors was linear and was (a) more accurate than the average clinical judge's accuracy and (b) also more accurate any of the 29 clinicians'

individual diagnoses. A concern, somewhat vitiated by the large number of subjects in the study, is that the mechanical prediction was not cross-validated, hence overestimating to some degree the mechanical accuracy of the true population model. Famously, very little predictive accuracy was lost by using, instead of optimized regression weights, equally weighted raw scores from four MMPI scales. These comparisons, then, support the conclusion that simple linear data combinations perform better than clinical judgments. Any mechanical prediction system that is at least as accurate as linear models would be expected to outperform clinician judgments.

Reviews: Narratives and Box Scores

Meehl (1954) simplified the prediction process used by both clinical and mechanical predictions to two stages: (a) data measurement (after the criterion construct has been defined and measures for all predictor constructs have been found or created and then measured); and (b) data combination joining measures of all or some predictor constructs to predict the criterion. Meehl presented in one chapter of his famous book a narrative review of prediction studies, with accompanying box score. Rough equality of clinical and mechanical prediction accuracies was considered sufficient cause to count a study as being promechanical. This is due to the fact that mechanical prediction is almost always cheaper than skilled clinician time, but it implicitly redefines the question being answered by comparison studies.

In the original box-score review, Meehl (1954) found 20 studies on point: The criterion for reviewing a study was that a prima facie fair comparison could be made between clinical and mechanical prediction methods. In all studies, mechanical prediction was approximately as accurate as, or more accurate than, clinical prediction. Although this "disturbing little book," as Meehl (1986) called it, devoted only a single chapter to the tabulation of study results, many important conceptual issues were addressed as well, with numerous arguments for and against clinical prediction being dissected.

Meehl (1954) indicated that before reading the studies, he was undecided, neither proclinical nor promechanical. In this vein, Meehl was moved by reading Lindzey's (1965) article, "Seer Over Sign" to write a response (Meehl, 1965), "Seer Over Sign: The First Good Example," which, as the title indicates, contains an evaluative review of Lindzey, culminating in the conclusion that Lindzey's clinicians did better than the mechanical predictions, unlike any of the 20 studies covered by Meehl (1954). Goldberg (1968), in "Seer Over Sign: The First 'Good' Example?" pointed out that Meehl (1965) had missed an error made by Lindzey in evaluating the degrees of freedom for crucial chi-square accuracy statistic. Correct evaluation led to the conclusion that clinical prediction accuracy was not significantly higher than that of mechanical prediction, contrary to Meehl's (1965) conclusion.

The first study reviewing investigations in the post-Meehl book era was by Holt (1958; expanded and with improved arguments in Holt, 1970). Holt did not offer a comprehensive survey of the literature, confining himself to conceptual analysis of methodology for a few key studies.

Another, very often cited, review was that by Sawyer (1966). Sawyer gave a box score across both clinical versus mechanical prediction and objective versus nonobjective data gathering (measurement), based on significance tests at the $p < .05$ level, but he did not follow this rule for all comparisons, detracting from the internal validity of his review. Sawyer had a polychotomous system for classifying 45 studies as to data gathering procedure (measurement, four values) and as to data combination procedure (two levels of the factor: clinical vs. statistical). Further classifying studies by type of accuracy statistic, (e.g., *r* vs. hit rate), Sawyer in this manner avoided solving the statistical problems dealt with by Grove et al. (2000). However, the type of accuracy statistic is irrelevant to the mechanical versus clinical comparison, and so Sawyer's treatment of the data represents added noise vitiating all comparisons.

Holt (1986) gave an extremely detailed and pointedly negative critique of Sawyer. The main points of Holt's evaluation are as follows. First, he stated that Sawyer made numerous errors in collating studies and classifying them into the big contingency table that was the Sawyer review's major output. Second, Holt argued that Sawyer accepted studies with poor

methodology, failed to make some important distinctions, and assumed that unmeasured variables did not bias studies for or against clinical prediction. Third, according to Holt, even presuming Sawyer's aim to classify studies as to data gathering and data combination was sensible, it is important that Sawyer misclassified procedures in a number of studies. Finally, Holt insisted that Sawyer's conclusions were influenced by numerous subtle but potent systematic biases against clinical judgment, so that Sawyer's overall conclusions were faulty.

Holt weighed in again in 1978, regarding his conception of the roots of clinical judgment. He held that a strong conclusion could not be drawn because of the paucity of well-designed studies. The Meehl (1954) conclusion that mechanical data combination was never materially less accurate than clinical data combination was strongly rejected by Holt. Holt pointed out that one would presumably like to compare mechanical prediction to clinical judgment at its best, not at its average value in daily life in the clinic. Holt argued that many studies, like Sarbin's (1943), where college GPA was predicted on the basis of high school rank and achievement testing, put the clinician at a disadvantage. He argued, first, that there were measures that no sensible clinician would use in clinical judgment, and there were studies describing predictions that a wise clinician would not undertake. Second, he pointed out that study designs were usually not arranged to control variation related to (a) choice of criterion variable, (b) methods of measuring a chosen construct, and (c) making good quality measurements. Hence, the numerous comparison studies with less-than-impeccable methodology should be ruled inadmissible to answer the question before us. Finally, Holt believed Meehl's, Sarbin's, and others' formulations of the problem involved a flawed question: Which data combination method is most accurate, averaging across inexperienced or inexpert clinicians working with unfamiliar variables and comparing clinical to mechanical accuracy? Many studies admittedly asked clinicians to use the variables picked out by the mechanical prediction algorithm. They do this for the obvious reason that it removes a potential confounder: amounts and types of data used for prediction. It is a question of enhancing internal validity, at the expense of external validity. Contrary to Holt's assertions, many other studies let clinicians use all available data in making predictions, with the mechanical predictions being based on a proper subset of the predictors used by clinicians (e.g., Blenkner, 1954). Holt argued in the affirmative part of his analysis that there was no advantage for mechanical data combination, when compared with a well-trained, seasoned, and expert clinician; however, this conclusion could not be strongly held because of the paucity of "good" studies (the same reason Holt gave to reject superiority of mechanical prediction).

Sarbin (1986) gave what we term *Sarbin*$_2$'s position, a quite different perspective from that presented in the 1940s, termed *Sarbin*$_1$'s position (e.g., Sarbin, 1944). The three essential points of Sarbin$_1$'s position were as follows: (a) A frequentist view of probability was relied on; (b) single-case probabilities were defined and measured by relative frequencies, even though Reichenbach held that single-case probabilities are meaningless; and (c) data combination by clinical versus mechanical predictions are wanted and can be obtained from studies like that of Sarbin (1943), which controls the "type of predictor data" while contrasting two methods of data combination. Sarbin$_2$ (1986) advocated a radically altered view. The four main points of the new theory are as follows: (a) The clinician reasons in a "narrative" mode of thinking that is fundamentally different from statistical prediction. (b) He states that mechanical prediction is validated by correspondence between facts and statements (Tarski, 1933). (c) Clinical predictions, on the other hand, are fundamentally different, in genesis and validation framework, from mechanical predictions. Clinical predictions do not aim at the same target as mechanical ones, so no legitimate comparison between the two can be made. Mechanical predictions try explicitly to maximize accuracy (or utility, in a decision-theoretic analysis), whereas clinicians construct narratives telling how a client comes to have a certain outcome, on the basis of all available information. Clinical "predictions" are validated by considering their coherence. (d) Because the coherence theory of truth applies to clinician reasoning, there are many clinical tasks in which a measure of

satisfactoriness for clinical predictions is incommensurable with the measure for mechanical prediction. Note that $Sarbin_2$'s view is relegated almost entirely to psychological decision-making and is most attractive in that arena. In medical decision-making, for example, where treatment operates on pathophysiology, it makes less sense to construct a narrative about the patient than it does to accurately identify his or her diagnosis and make a prediction about amelioration in presence of a diverse number of available treatments. $Sarbin_2$'s views make less sense in, say, a gambling scenario where decisions are made to obtain the most money possible.

At the very end of his 1986 remarks, Sarbin admitted the point that given his reconceptualization of clinical reasoning, the question changes from "Which data combination method is more accurate?" on average to "Which conception of truth—correspondence or coherence—is best applied to which specific clinical tasks and situations?" Without a set of measures of the degree to which correspondence theory speaks for the truth of statements such as "Actuarial prediction is as satisfactory as, or more satisfactory than, clinical prediction," this is no longer an empirical question, whereas $Sarbin_1$'s problem was empirically addressed by various studies. As such, for $Sarbin_2$ a clinical versus mechanical prediction must be either a deductive logic problem or a semantic one.

Hence, even with the best-designed large-sample meta-analysis of the question before us, one cannot under this conception comprehensibly compare clinical predictions to mechanical predictions. It would make any narrative review of studies (e.g., Meehl, 1954) yielding a box score and any meta-analysis completely pointless. However, even if $Sarbin_2$ were completely correct, one could still render a narrative (clinical) "prediction" into something to which the concept of correspondence-based truth would relate. We also point out something not mentioned by $Sarbin_2$; namely, that a narrative can be read by a trained rater and turned into a judgment about an outcome, to which the concept of an accuracy statistic (based on the correspondence theory of truth) applies, even though the narrative basis of the rating might well be congenial to the narrative form of clinical judgment.

Finally, Dawes, Faust, and Meehl (1989) gave a box score based on trichotomization of preliminary quantitative statistics from the Grove et al. (2000) meta-analysis and, hence, not reported here. Trichotomization refers to three study outcomes: (a) clinical prediction superior; (b) clinical prediction approximately equal to mechanical; and (c) mechanical prediction superior. The same procedure as in Meehl (1954) was followed; namely, that Type 2 study outcomes are counted together with Type 3 study outcomes. This supported the conclusions "Actuarial prediction is always at least approximately as accurate as clinical prediction," and "Actuarial prediction is often materially more accurate than clinical prediction."

Literature Reviews: Meta-Analyses

To the best of our knowledge, Grove et al. (2000) published the first meta-analysis of articles on clinical versus mechanical prediction. One hundred thirty-six studies having prediction tasks related to human health and behavior were coded and used, giving multiple effect sizes (ESs) for many studies (most notably, Goldberg, 1965, which has nearly as many ESs as all the other studies put together). The weighted median ES for each study was used to create a study ES. The distribution of these study ESs then dictated our conclusions. Please bear in mind that averaging ESs within a study to obtain a study ES fails to take account of the way in which drawing ESs from the same study creates correlated ESs. This is not a data analysis strategy one would ordinarily favor, but in this situation there was no real choice. The implication is that the standard deviation of ES figures will be misestimated, with homogeneity (Q) statistics biased downward. Unlike Ægisdóttir et al. (2006), we calculated Q statistics but we did not remove outliers to produce a nonsignificant Q statistic, as the other investigators did.

The primary result of Grove et al. (2000) was that mechanical prediction was about 10% more accurate than clinical prediction. The effect held whether experienced or novice clinicians were considered. It held whether one considered educational, financial, forensic, medical, clinical psychology, or personality psychology outcomes.

Ægisdóttir et al. (2006) conducted a long-running study of clinical judgment. This large-scale study started with searching the literature with 207 search terms, then encoded 1,135 published and unpublished studies. After limiting the scope of attention to studies making direct comparisons of clinical and mechanical prediction and to predictands that clinical psychologists routinely assess, they reduced their massive database to 69 studies. If more than one ES was found in a study with different study designs, all were encoded. Two of the studies, one of them being the Goldberg (1965) study, were included in some analyses and excluded in others. This is because these two studies had so many ESs per study (81 in two studies) that they threatened to swamp the results. They encoded multiple ESs from a single study for analysis if different ESs were generated with different study design features. If both cross-validated and unvalidated statistical prediction rules were present, only the cross-validated ones were included. These investigators ended up with 173 ESs, 92 excluding the two studies. Next, outliers were detected statistically by significant *Q* statistic (a homogeneity test), and for some analyses they were excluded.

The main result was that mechanical prediction was 0.12 standard deviations more accurate than clinical prediction. Inclusion of outliers did not change the result notably. Twenty-five ESs were more than 0.1, five were less than −0.1, and the balance were in between. This is very similar to what Grove et al. (2000) found. Moderator effects were identified: type of prediction task (prediction of prognosis, criminal offense, and academic achievement were more predictable), setting (i.e., clinicians making predictions in their customary setting vs. a different setting with, surprisingly, the direction of the effect favoring the algorithm more when clinicians were operating in their usual setting), and linear formulas faring better, in comparison with clinical judgment, than logical rule sets.

REASONS WHY MECHANICAL PREDICTION IS SELDOM USED

Another important aspect of Meehl's (1954; Grove & Meehl, 1996) work was to theorize about reasons why clinicians apparently do not use mechanical prediction very often. These were the potential reasons he listed in 1954 and 1997:

1. ignorance of the controversy;
2. fear of technological unemployment;
3. protection of self-concept;
4. theoretical identifications that do not take into account prediction;
5. dehumanizing flavor of using mechanical procedures instead of the clinical approach;
6. mistaken conceptions of the ethics of such predictions; and
7. computer phobia.

The list was not composed on the basis of empirical studies of the clinician's approach to combining data to generate predictions, and it is not even clear that Meehl informally surveyed his clinician acquaintances.

Vrieze and Grove's (2009) survey of 491 U.S. clinicians found that 40% of clinician respondents stated they did not use mechanical prediction because there was none available for their particular prediction problems. Thirty-six percent stated they were familiar enough with mechanical prediction methods to be comfortable using them. Thirty-two percent stated that mechanical predictions are not as accurate as clinical predictions. Thirty-two percent stated that mechanical predictions cannot possibly account for all factors that influence a prediction. Other reasons were also endorsed but with less frequency.

For the interested reader, mechanical prediction tools can be quite easily constructed. If there exist predictors known to be valid for some prediction, then regression formulas can be constructed immediately, using either equal weights (Wainer, 1976) or correlation weights (Waller & Jones, 2010). Large validation and cross-validation sets are simply not necessary. Equal weights and correlation weights are immediately applicable in clinical settings, as long as the predictor variable information is available (e.g., in the published literature) or gathered. There is also a literature on "fast and frugal" reasoning, where, for example, only the known best predictor is used, and all other information is thrown away (Gigerenzer & Goldstein, 1996).

In a compelling and paradoxical study, Goldberg (1970) found that regressions constructed to predict clinicians' predictions (i.e., not the outcome itself, but what the clinicians predicted the outcome to be) fared better than the same clinician's own predictions when making novel predictions of the same kind. Thus, anywhere clinical predictions are made, a formula can be constructed to predict those predictions. One will expect the resulting formula to be more accurate than the clinicians on whom it was constructed. An explanation for Goldberg's (1970) result is that clinicians are unreliable. Their predictions differ from day to day, depending on contextual factors (e.g., how much caffeine they have had) that influence predictions, have no validity, and add noise to the prediction. The model of the clinical prediction filters this noise to some extent and has improved cross-validated accuracy.

The literature on clinical judgment heuristics, biases, and intuition is vast and not covered here to any appreciable effect. Seminal work on this topic was conducted by Kahneman, Slovic, and Tversky (1982). Clinical judgment biases are covered comprehensively by Garb (1999), and we refer the interested reader there.

SUMMARY

Across varying criterion variables, types of judges, and predictor variables, the research nearly uniformly shows that mechanical prediction is either more accurate than or is approximately as valid as clinical prediction, according to two independent meta-analyses. Meta-analyses show nearly identical average effect sizes, about 9% to 10% of a standard deviation. Algorithms to systemize clinical judgment (i.e., building models of judges' clinical reasoning and using the model to make mechanical predictions) have been investigated. Unfortunately, studies show that this still produces predictions that prove to be more valid for mechanical prediction than output of average, or even the most accurate, clinical predictions. Finally, only a few moderator variables influence effect sizes for clinical versus mechanical prediction significantly, and then they do not do so strongly. That is to say, outcomes of clinical–mechanical prediction studies have quite heterogeneous outcomes, and we do not know why. Future research needs to bring order to study outcomes, as this information would be invaluable for improving predictive accuracy.

This chapter has focused on an informal survey of the literature, including some single studies, narrative reviews, and meta-analyses. Holt (1978) may have argued that clinical prediction is more accurate than mechanical prediction for some experts and in some situations. This result does not exist in the literature. For mental health and medical outcomes, there are no high versus low expertise differences and no criterion variables behaving differently from other criterion variables in study outcomes, such as predicting personality or vocational outcomes (Grove et al., 2000).

It may very well be that there are situations in which clinical prediction is more accurate or, in the presence of cost information, less costly to patients and the clinic overhead than mechanical prediction. That result is not to be found reliably in the literature. We conclude that, whereas clinical prediction may fare better, there is no evidence for that conclusion, and proclinical armchair arguments only count so much. Perhaps the most biting complaint by clinicians is that mechanical prediction schemes are too difficult to understand and that many times none exist for a particular prediction problem (Vrieze & Grove, 2009). Both complaints are easily remedied. Clinicians can be trained, and new mechanical schemes can be constructed quickly and cheaply using equal/correlation weights, fast and frugal reasoning, or models of clinician predictions.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382. doi:10.1177/0011000005285875

American Psychiatric Association. (2000). *Diagnostic and statistical manual for mental disorders* (4th ed., text revision). Washington, DC: Author.

Blenkner, M. (1954). Predictive factors in the initial interview in family casework. *Social Service Review, 28*, 65–73. doi:10.1086/639511

Dahlstrom, W. G., Welsch, G. S., & Dahlstrom, L. (1972). *An MMPI handbook.*: *Vol. 1. Clinical interpretation.* Minneapolis: University of Minnesota Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674. doi:10.1126/science.2648573

Garb, H. N. (1999). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669. doi:10.1037/0033-295X.103.4.650

Goldberg, L. R. (1965). Diagnosticians or diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs, 79*(9, Whole No. 602).

Goldberg, L. R. (1968). Seer over sign: The first "good" example? *Journal of Experimental Research in Personality, 3*, 168–171.

Goldberg L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 422–432.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323. doi:10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30. doi:10.1037/1040-3590.12.1.19

Hájek, A. (2007). The reference class problem is your problem too. *Synthese, 156*, 563–585. doi:10.1007/s11229-006-9138-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1–12. doi:10.1037/h0041045

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25*, 337–349. doi:10.1037/h0029481

Holt, R. R. (1978). *Methods in clinical psychology: Vol. 2. Prediction and research*. New York, NY: Plenum Press.

Holt, R. R. (1986). Clinical and statistical prediction: A retrospective and would-be integrative perspective. *Journal of Personality Assessment, 50*, 376–386. doi:10.1207/s15327752jpa5003_7

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.

Korman, A. K. (1968). The prediction of managerial performance: A review. *Personnel Psychology, 21*, 295–322. doi:10.1111/j.1744-6570.1968.tb02032.x

Kyburg, H. (1978). Subjective probability: Criticisms, reflections, and problems. *Journal of Philosophical Logic, 7*, 157–180.

Lindzey, G. (1965). Seer over sign. *Journal of Experimental Research in Personality, 1*, 17–26.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. doi:10.1037/11281-000

Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality, 1*, 27–32.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375. doi:10.1207/s15327752jpa5003_6

Pollock, J. (1990). *Nomic probability and the foundations of induction* (pp. 234–254). Cambridge, England: Oxford University Press.

Pollock, J. (2007). The *Y*-function. In G. Wheeler & B. Harper (Eds.), *Probability and evidence*. Cambridge, England: King's College Publications.

Reichenbach, H. (1949). *The theory of probability*. Berkeley: University of California Press.

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*, 593–602. doi:10.1086/219248

Sarbin, T. R. (1944). The logic of prediction in psychology. *Psychological Review, 51*, 210–228. doi:10.1037/h0057400

Sarbin, T. R. (1986). Prediction and clinical inference: Forty years later. *Journal of Personality Assessment, 50*, 362–369.

Sawyer, J. (1966). Measurement and prediction: Clinical and statistical. *Psychological Bulletin*, *66*, 178–200. doi:10.1037/h0023624

Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry, 116*, 129–144.

Tarski, A. (1933). *The concept of truth in the languages of the deductive sciences* (No. 34) [in Polish]. Warsaw, Poland: Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych.

von Mises, R. (1957). *Probability, statistics and truth* (2nd revised English ed., prepared by Hilda Geiringer). New York, NY: Macmillan.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40*, 525–531. doi:10.1037/a0014693

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213–217. doi:10.1037/0033-2909.83.2.213

Waller, N. G., & Jones, J. A. (2010). Correlation weights in multiple regression. *Psychometrika, 75*, 58–69. doi:10.1007/s11336-009-9127-y

CHAPTER 5

EDUCATION AND TRAINING IN ASSESSMENT FOR PROFESSIONAL PSYCHOLOGY: ENGAGING THE "RELUCTANT STUDENT"

Beth E. Haverkamp

Readers approaching this chapter are likely to be positively disposed toward the use of standardized tests in psychology and knowledgeable about psychometrics and the integration of test data in psychological practice. That is not the case for many students. Many assessment instructors have entertained the question, "Is this course required? I don't plan to use tests with my clients." These are the students who can make teaching assessment a challenge, and it is a challenge that can engage the instructor's own professional identity as much as it does students' attitudes and preconceptions.

This chapter is concerned with the project of teaching assessment to graduate students in psychology and grows out of more than 20 years of faculty experience in designing and delivering graduate courses in measurement and the use of standardized tests. The consideration of what topics to address began with the question, "What do I try to accomplish in my own teaching?" and the first words that came to mind were "interest, motivation, and engagement." These classes are designed to provide a strong knowledge base across the range of competencies required for effective test use, but readers will share the perception that the fundamental challenge in teaching assessment has little to do with students' capacity for knowledge acquisition. Typically, such classes are filled with high-achieving young professionals who have navigated their way to graduate school in psychology but often include a subgroup who approach the use of tests with ambivalence, suspicion, and even aversion. In subsequent sections, this chapter presents the argument that creating a successful environment for learning requires that faculty engage directly with student attitudes and values as well as provide substantive content on tests and measurement.

The ambivalence with which some students approach assessment coursework is paradoxical given their choice of a career in psychology and the historical and contemporary role of psychological testing within the field. There is broad recognition that assessment continues to be widely used by psychologists in key societal and individual decisions (e.g., evaluation of prison inmates for parole; student access to special assistance and education; child custody decisions, career choice, and treatment planning). Krishnamurthy et al. (2004) have reminded us that assessment has long been a defining characteristic of applied psychology, that the field continues to be the primary locus of assessment knowledge and use, and that tests and assessment are integrated in all areas of applied psychology practice.

However, several worrisome signs have emerged and, in conversations about the status of assessment practice within the field, the teaching of assessment is receiving renewed attention (e.g. Krishnamurthy et al., 2004). Despite decades of research demonstrating that standardized assessment provides a trustworthy guide to individual and organizational decision making as well as more recent documentation of its effectiveness as a therapeutic intervention (Finn & Tonsager, 1992; Poston & Hanson, 2010), there has been a significant drop in psychologists' use of standardized assessment (Eisman et al., 1998)

DOI: 10.1037/14048-005
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

and in the level of assessment training provided in graduate programs (Belter & Piotrowski, 2001; Stedman, Hatch, & Schoenfeld, 2001). Although several of these authors (e.g., Eisman et al., 1998) have cited the constraints of managed care and third-party reimbursement as likely culprits in these changes, there has been little sustained examination of the level of interest that students bring to this area of professional practice, although it may be another influential factor. A 1980 survey of Canadian undergraduate career preferences in psychology (Babarik, 1980) found that of 1,360 students surveyed, only two chose evaluation and measurement as their most preferred area of specialization.

Of equal concern, some indicators raise questions about whether test use is conducted competently, effectively, and ethically. As examples, Koocher and Keith-Spiegel (2008) have argued that a failure to apply basic psychometric knowledge often underlies ethics complaints related to assessment; Curry and Hanson (2010) found that only 35% of clinical, counseling, and school psychologists provide verbal test feedback consistently; and a review by Alfonso and Pratt (1997) found that both graduate and professional psychologists make frequent errors in the administration and scoring of cognitive ability measures.

These events are occurring in the midst of debate within the teaching profession on the utility of large-scale benchmarking and achievement assessments as well as the proliferation of unvalidated self-help "measures" on the Internet (LoBello & Zachar, 2007). Societal attitudes toward assessment, as reflected in contemporary news stories, appear to be more negative than positive, and it is fair to assume that some students are influenced by attitudes within the society at large.

The field needs to consider these societal and professional trends and to incorporate that understanding in preparing the next generation of psychologists as competent and ethical users of standardized assessment. Without question, the content domain associated with teaching assessment, as broadly defined, is diverse and can range from a knowledge of measurement, psychometrics, and familiarity with standardized tests to the less structured process of conducting effective intake interviews, to using psychodrama techniques such as sculpting as a means of assessing family dynamics. Because empirically based forms of assessment are those most likely to meet with student resistance, this chapter is limited to issues and concerns related to introducing students to standardized assessment.

Before addressing questions of what to teach or how to teach, it is critical to consider *who* is in our classrooms. This question can be considered in two areas: First, it is important to acknowledge the diversity of psychology students who take assessment courses. There can be significant variation in student background across different specializations within applied psychology, as students choose career paths consistent with their interests. Furthermore, different work settings call for variable emphasis on different elements of the assessment process. For example, areas that emphasize use of tests for high-stakes decisions (e.g., personnel, corrections) are likely to place great emphasis on issues of accuracy and validity and may have higher expectations that students learn a measure's psychometric characteristics in detail (e.g., characteristic scale intercorrelations or configurations that typify distinct populations). On the other hand, specializations that emphasize the use of test data as an aid to client exploration (e.g., vocational counseling) or that depend on strong rapport to elicit maximum performance (e.g., cognitive ability testing with children) may give greater attention to the relational aspects of test interpretation.

Students across specializations need to gain competence in all aspects of the assessment process, and one of the challenges in designing assessment curricula is managing and accessing the large body of knowledge students must acquire to become competent assessment professionals; a new instructor can feel uncertain about what areas are most important to cover within a single term. Ideally, a single course will serve a defined purpose in a staged process of learning and be integrated with student experience in clinics and practica. Given the goal of helping students achieve mastery across the range of assessment competencies, it may be important for instructors and supervisors to give particular attention to areas that fall outside the typical area of emphasis within their specialization. For example, students in

areas with a predominant analytic focus may need additional instruction or supervision to develop the rapport skills necessary for effective administration and interpretation; those in areas that emphasize therapeutic relationships may need additional work on the psychometric knowledge necessary for appropriate test selection and interpretation. The basis for this recommendation addresses the potential risk of test misuse associated with acting in areas of testing competence that are underdeveloped.

Here, several key resources for assessment instruction that have appeared in the past decade are described, associated with the movement to define competencies in professional psychology. Following that description are topics that are absent from these lists or, at a minimum, merit greater attention.

COMPETENCIES IN ASSESSMENT

Psychologists tasked with assessment course development and instruction have gained a wealth of new resources in the past decade. Most noteworthy among these, in addition to the present handbook, is the report produced by the Assessment of Competency Benchmarks Work Group (Fouad et al., 2009), convened in 2005 by the American Psychological Association (APA) Board of Educational Affairs (BEA), in collaboration with the Council of Chairs of Training Councils (CCTC). Building on the work of the 2002 "Competencies Conference: Future Directions in Education and Credentialing in Professional Psychology," (Kaslow, 2004; Kaslow et al., 2004) and the competency cube model developed by Rodolfa et al. (2005), the Benchmarks Work Group identified the area of assessment as one of the core functional competencies required for effective psychological practice.

Of documents emerging from the competencies movement, those of particular importance for instructors are the final report of the APA BEA Task Force on Assessment of Competence in Professional Psychology (Fouad et al, 2009), the report of the Psychological Assessment Work Group at the 2002 Competencies Conference (Krishnamurthy et al., 2004), and the Competency Assessment Toolkit for Professional Psychology (Kaslow et al., 2009), produced by the aforementioned Benchmarks Work Group. Together, these documents provide an invaluable description of the knowledge and behaviors that are the intended outcomes of assessment instruction across the curriculum. The Benchmarks Work Group report (Fouad et al., 2009) covers the range of assessment practices, with sections on measurement and psychometrics, evaluation methods, application of methods, diagnosis, conceptualization and recommendations, and communication of findings. The report enumerates the competencies, with behavioral anchors appropriate to various levels of trainee development (e.g., readiness for practicum, internship, and entry to practice) and with a particular focus on preparation for health service practice. In addition, the group has produced a range of supporting materials for evaluation of competencies, available on the APA Education and Training Web site.

The report of the 2002 "Competencies Conference" work group on psychological assessment (Krishnamurthy et al., 2004) became a source document for the 2007 competency benchmark project and identified eight core competencies in the area of psychological assessment, which are largely incorporated in Fouad et al. (2009). In discussing the evaluation of competencies, the work group was guided by a model used in industrial and organizational psychology: knowledge, skills, abilities, and other characteristics (KSAO). As Krishnamurthy et al. (2004) noted, the KSAO framework can be used for planning and curriculum design as well as evaluation of individual performance; they offered the following definitions of the four KSAO components:

> Knowledge refers to the psychometric and theoretical information acquired through coursework; Skills refers to proficiency in different methods of assessment (e.g., test administration, scoring and interpretation; interviewing; observations) and communication of assessment findings; Abilities include rapport building, critical and integrative thinking, and psychological mindedness; Other Characteristics could include attitudes and values such as respect for the person of the client and appreciation of diversity, and a variety of facilitative capacities

> such as precision/accuracy, attention to detail, and good communication skills. (p. 734)

The Assessment of Competency Benchmarks Work Group incorporated the KSAO rubric in its work, where each of 12 core competencies was classified as either foundational or functional. Assessment is designated as one of the six functional competency domains (with, e.g., intervention and research/evaluation); the six foundational competencies (reflection/self-assessment, scientific knowledge and methods, relationships, ethical and legal standards, individual and cultural diversity, and interdisciplinary systems) are described as "the knowledge, skills, attitudes and values that serve as the foundation for the functions that a psychologist is expected to perform" (Fouad et al., 2009, p. S6). It is worth noting, however, that the resulting benchmarks are not intended to represent the intersection of the six foundational competencies with the functional domain of assessment (or others); there is no benchmark, for example, representing the intersection of relationships and assessment, although a behavioral anchor for entry to practice is "provides meaningful, understandable and useful feedback that is responsive to client need" (Fouad et al., 2009, p. S17).

The benchmark project and the KSAO framework are discussed in some detail because this rubric, although invaluable as a map for the desired outcomes in assessment instruction, also reveals neglected areas that have potential to enhance the effectiveness of assessment education and training. Specifically, there is value in examining which aspects of the KSAO model are not typically represented in assessment courses or statements of desirable assessment competencies.

The KSAO framework can be considered in light of the three questions posed earlier: what is taught, how it is taught, and who is taught. The great majority of the competencies identified and the overwhelming content of traditional assessment and testing courses are most relevant to the question of "what is taught" and are focused on the knowledge and skill domains. However, given Clemence and Handler's (2001) finding that most students enter internship without basic, requisite assessment skills, one has to suspect that even the skills domain is neglected in teaching and assigned to the internship experience. The same authors, in their survey of 382 psychology internship sites, found that fully 56% of the sites had to provide interns with basic assessment training. Furthermore, Curry and Hanson (2010) found that one third of clinical, counseling, and school psychologists surveyed reported that their graduate training (coursework, practica, and internship) were "of little to no help in preparing them to provide feedback" (p. 327) on client test results. Krishnamurthy et al. (2004) reported that doctoral programs are devoting fewer course credits to assessment instruction and preparation; if a large amount of content is squeezed into a smaller container, it becomes even more important to consider carefully what content is most essential to convey. The tables of contents for the three volumes of this handbook are a rich menu of options that can inform development of a course syllabus, particularly in specific areas of assessment knowledge (the K in KSAO). Additional chapters describe the assessment process, the skill of writing reports, and communication of test results; these are important as resources for addressing skills (the S in KSAO).

The remainder of this chapter focuses on the KSAO domains of abilities and other characteristics (the A and the O in KSAO). Both have been generally disregarded in assessment education and training, although they are defined as competencies necessary for effective practice. Direct attention to abilities and other characteristics can also enhance instructional effectiveness. As a reminder, the abilities domain encompasses relational and critical thinking abilities; the other characteristics domain is centrally concerned with attitudes and values. The present discussion begins with the O in KSAO because student receptivity to a wide range of content is often directly related to an instructor's success in creating a receptive climate for learning.

BARRIERS TO STUDENT ENGAGEMENT: THE "O" OF COMPETENCE, ATTITUDES, AND VALUES

A consequential issue in considering "who is taught" concerns a group that can be described as *reluctant*

students, those who enter such classes with a set of beliefs and/or values that make them ambivalent, suspicious, or even hostile to the whole enterprise of standardized assessment. Students who are at greatest risk for not engaging with assessment instruction may hold a range of inaccurate perceptions and biases. The following phrases may sound familiar to assessment instructors: "Tests put people in boxes and label them," "Testing requires me to take on an expert role and I want to be collaborative with my clients," "Someone said all tests have error, so why rely on them?" These comments may be more frequent in therapeutically oriented specializations such as counseling and clinical psychology, but reports from colleagues in school psychology and industrial–organizational psychology suggest that they can appear there as well. Although these stereotypes are inaccurate, they do constitute a barrier to student engagement.

The more positive pole of the beliefs, attitudes, and values reflected in these stereotypic comments can be rephrased as, "I want to treat clients as individuals," "I pursue collaborative relationships," "I view clients as experts on their own experience," and "I don't believe objective description is really possible." It is important to note that these assertions echo some of the core tenets of humanistic and postmodern, constructivist approaches to psychology that have become increasingly influential in therapeutic psychology (e.g., Aschieri, Finn, & Bevilacqua, 2010; Neimeyer, 1995). Students influenced by postmodern perspectives on philosophy of science and psychotherapy are likely to have a stance toward psychological practice that emphasizes relational, contextual information and a more emic than etic perspective. This group is also more likely to question the realist, objectivist stance of positivism and to be more aligned with understanding individuals in context, viewing reality and meaning as constructed within relationship, and to have greater trust in the processes of intuition and "meaning making."

The relevance of these trends for teaching assessment is that they may contribute to an increased (perceived) bifurcation between the traditions of test use and the values and assumptions of students who want to pursue a therapeutic career, with respect to a stance toward knowledge (epistemology) and the psychologists' role (axiology). The inherent epistemology of tests is rationalist and reductionistic—that is a large part of their utility—and the domain of psychometric knowledge is deeply grounded in a positivist/postpositivist philosophy of science, which assumes a knowable, objective reality and positions the scientist as a neutral, rather distant observer. The positivist, empirical/realist philosophy of science characterized the training received by anyone who graduated in the 20th century and continues to be the dominant worldview in most graduate departments, particularly in the more statistically grounded specializations of measurement and assessment. In contrast, as noted, an increasing number of graduate students are drawn to the tenets of a postmodern, constructivist philosophy of science and to therapeutic approaches that view "meaning" as intersubjective, or jointly constructed within social discourse. Such perspectives challenge notions of objectivity, so central to the use of tests, and prioritize psychologist–client collaboration on tasks and goals (e.g., Horvath & Bedi, 2002). An "expert" stance is roundly criticized; Corey (2009), whose texts are widely used in training of psychotherapists, noted that postmodern therapists "adopt a stance characterized by respectful curiosity . . . the client is the expert when it comes to what he or she wants in life" (p. 390).

Instructors may have a tendency to conclude that reluctant students simply have insufficient interest in the data-based domain of testing and hold stronger "people" interests, congruent with the formulation of "data/ideas–people/things" (Prediger, 1982; Prediger & Swaney, 2004). This is likely to be an oversimplification of the gap between postmodern students and their positivist/postpositivist assessment instructors. The perspective one holds on a philosophy of science can engage one's fundamental worldview and belief system in that it is concerned with what one believes can be known, how things can be known, and what is valued.

A final strand within therapeutic psychology that may be perceived as incompatible with traditional approaches to testing is the increased attention to the role of culture and diversity. Mintz et al. (2009), writing on the subject of diversity and values, argued that the postmodern trends in both research

and practice have created the awareness that "knowing" is embedded in culture and context, that values are always present in practice, and that one must acknowledge a plurality of perspectives over a single, knowable reality.

This is not the first time that questions have been raised about the compatibility of therapist attitudes and values and standardized assessment. In 1990, Watkins and Campbell described a renewed focus on "the person" in assessment and a concern for "humanizing assessment" (p. 193). Although welcoming many aspects of this trend, the authors also noted, "Exactly how this (humanistic) perspective will conflict or co-exist peacefully with the focal assessment movement . . . is unclear at this time" (p. 194). Although not particularly conflictual, a lessening of student interest and engagement has been observed in assessment, particularly among those choosing private practice careers in clinical and counseling psychology. This parallels the finding of Camara, Nathan, and Puente's (2000) survey of applied psychologists, which documented a decline in test use.

With this state of affairs, explicit attention to philosophy of science issues, as part of classroom discussion, can advance Watkins and Campbell's (1990) hopes for peaceful coexistence. Cacioppo (2004) has noted that an instructor's explicit discussion of the philosophy of science is an effective way to deal with what he has termed student *entry biases* with respect to a discipline's implicit assumptions and boundary conditions. Of note, he did not dismiss the value of intuition, stating,

> To be clear, intuitions can foster or hinder theoretical progress in a scientific discipline. In personality and social psychology, the subject matter is so personal that many of the intuitions, prior beliefs and naïve theories people bring to the discipline are based on unsystematic experiences and observations. Our aim here is to encourage recognition of the power of intuitions in theory construction and hypothesis testing and to consider means by which naive intuitions might be evaluated and, as necessary, refined. (p. 115)

One mechanism for introducing the philosophy of science to assessment classes is to generate discussion by means of a heuristic, two-dimensional grid that represents the epistemological and axiological dimensions of test use (see Figure 5.1). The respective positions of the reluctant student and traditional

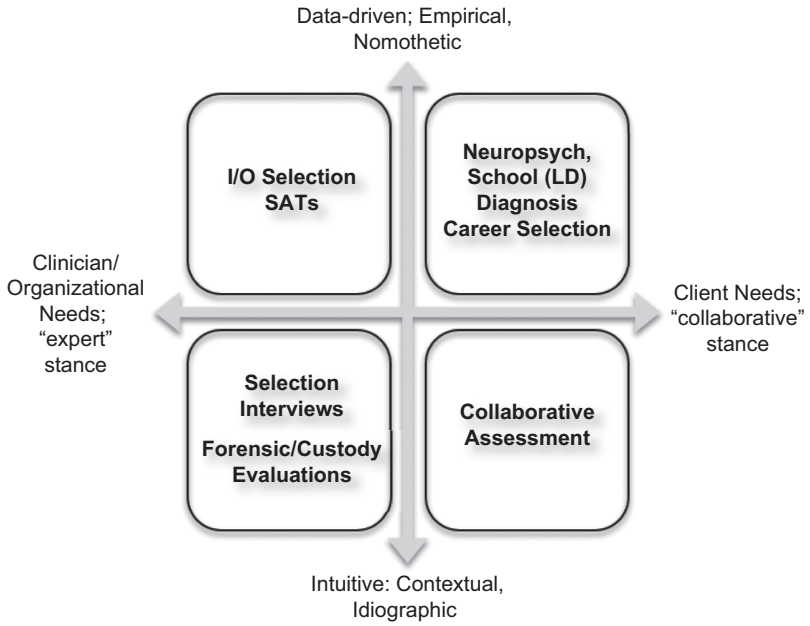


FIGURE 5.1. Epistemological and axiological dimensions of standardized assessment. I/O = industrial–organizational; LD = learning disability.

users of standardized testing can be represented in two-dimensional space describing the process of assessment, broadly defined; furthermore, if familiar types of assessment are mapped on the resulting quadrant, students with a more constructivist, postmodern, or client-centered orientation can readily identify areas of the traditional assessment landscape that create the greatest discomfort while also recognizing areas that are more compatible with their stance toward practice. The grid's heuristic value is to enhance understanding, acknowledge diverse perspectives, and increase engagement.

The quadrant is defined by two dimensions that describe important aspects of assessment and that can be roughly aligned with the positivist/postpositivist and postmodern philosophy of science traditions. The *x*-axis designates the purpose of assessment, with the associated question of whose needs are given priority, and corresponds to the values, or axiology, of the testing event. The *y*-axis indexes the basis for making inferences from the information gathered, or the epistemology that underlies various types of inferences.

The purpose of assessment carries an implicit assumption about whose needs or interests are primary in conducting standardized testing and can range from (a) exclusive priority to an organization's or psychologist's need for trustworthy information to (b) testing conducted primarily to serve a client's interest in exploration or self-assessment. The former is associated with a more traditional "expert" stance and is likely to be associated with greater vulnerability on the part of the client on the basis of the assumption that they may have less say in whether assessments are conducted. The latter form of assessment is more likely to include clients in the decision about whether to use standardized testing in service of their goals and would be associated with a more collaborative approach and less client vulnerability. This characterization is not new; in counseling psychology, there is a long-standing tradition of involving clients in the decision to conduct testing as well as in the interpretation and application of results (e.g., Duckworth, 1990; Tinsley & Bradley, 1986).

The *y*-axis, and the epistemology underlying the basis for inference from test results, is rarely discussed. One pole is defined by the traditional empirical, positivist perspective and can be characterized as a data-driven, nomothetic stance. The opposite pole is more closely aligned with idiographic, contextual, and intuitive approaches to inference and is more aligned with the emerging constructivist stance within psychology as well as with recent developments in applied psychology that direct attention to culture, diversity, and local knowledge. The *y*-axis could also be interpreted as relevant to whether one has a primary interest in drawing conclusions, where accuracy is paramount, or in developing hypotheses, where accuracy may still be valued but viewed as tentative.

It is informative to map some well-known forms of testing onto the quadrant and to consider the implications with respect to epistemology and axiology. Some students in clinical and counseling psychology—whose training asks them to embrace the field's call to attend to diversity; to develop therapeutic alliances based on collaboration in the tasks, bond, and goals of therapy; and to view all clients as distinct individuals—can hold an assumption that the lower right-hand corner is empty. A final section of the present chapter identifies an approach that can populate that quadrant; for now, this hypothetical description is offered to illustrate the sense of alienation that some students experience when they enter their required measurement and assessment courses. The utility of bringing the grid into classroom discussion is that by explicit consideration of philosophy of science perspectives, the instructor has a structure for respectful acknowledgment of diverse views, a language for discussing differences, and a vivid illustration of areas where students may position themselves as well as an opportunity to identify assessment approaches consistent with their approach to clients.

To summarize, a bifurcation appears to exist between the values, attitudes, and implicit philosophy of science held by growing numbers of students and the traditional stance and philosophy of science inherent in standardized assessment. The contrast and the resulting tension are genuine and not something that can be indoctrinated away by a deep immersion in the subject matter. Instead, instructors are encouraged to bring this conversation into the

classroom as a means of creating bridges between student perspectives and the field's valuing of standardized testing. As Levin (2008) noted,

> Knowledge by itself is not enough to change practice, since practices are social and therefore reinforced by many elements such as norms, cultures, and habits. Simply telling people about evidence and urging them to change what they do is clearly ineffective. (p. 8)

For instructors who recognize the reluctant students described earlier, what are the implications of this portrait for teaching courses on standardized assessment? As noted in Levin's (2008) remarks, a traditional approach to curriculum and instruction is unlikely to be sufficient to engage such students with the course objectives. A review of doctoral program requirements and textbooks in assessment suggests that most instruction follows a traditional model of coursework in measurement and psychometrics, followed by an additional course and practicum supervision in the use of tests, including exposure to a range of widely used measures. However, for a student who does not view standardized tests as particularly relevant to their future career practice, there is nothing in this model that will help them bridge this divide. This introduces a risk that material will be learned at a superficial level and that students will not achieve the level of integration between psychometrics, test characteristics, and test use that is essential to effective practice.

The challenge with such students is to facilitate engagement with course material, which has to begin with attention to how tests can be relevant to the work they plan to do. Attitudes and values are central to the challenges of establishing engagement and relevance and, as noted previously, are cited as elements of assessment competency, in the KSAO framework as "other characteristics," beyond knowledge, skills, and abilities. Instructors may not view their role as requiring them to target student attitudes and values, particularly in the more academic parts of the psychology curriculum, and the prospect of engaging in "attitude change" may make some uncomfortable. However, this is exactly the area that is typically overlooked, with a result that students may complete their training without having engaged in a thoughtful debate about the role of psychological testing.

For instructors to be successful in facilitating student engagement, it is useful to conceptualize this as another area of competence, within the same KSAO framework. It may be that knowledge and skills are a key focus of assessment courses because these are also the competencies that are strongest for most instructors. To facilitate students' exploration of their attitudes and values related to testing, it is incumbent on instructors to engage in self-assessment and self-reflection on their own "O" competencies, their attitudes and values. The two areas where this appears to be most important are one's philosophy of science and one's view of the role of an instructor. Some instructors may be unfamiliar with the challenges to positivism as well as alternate perspectives that have been advanced; others may be acquainted with postmodern approaches but find it difficult to accept the alternate stance on objectivity and multiple realities. Whatever one's position, being an effective instructor and being able to engage students in the course material requires an examination of one's own philosophy of science position, an honest acknowledgement of one's reactions and values, and a willingness to understand key elements of alternative perspectives.

Once instructors accept the idea that their role includes responsibility for facilitating student engagement and an examination of attitudes and values, they will find that existing psychological research suggests ways to proceed. In therapeutic psychology, counseling or clinical interventions have been conceptualized as a deliberate attitude change process (see Heppner & Claiborn, 1989; Strong, 1968) in service of a client's goals. From academic psychology, the social judgment theory (Sherif & Hovland, 1961; Sherif, Sherif, & Nebergall, 1965) concepts of latitude of acceptance and latitude of rejection are useful reminders for the presentation of material inconsistent with a person's current attitudes. Research on motivation is particularly relevant to the issue of student engagement, and one model that has been widely applied in educational contexts is self-determination theory (SDT; Ryan & Deci, 2000; Ryan, Kuhl, & Deci, 1997),

which offers a detailed description both for how motivation is linked to learning and the role instructors can play in facilitating that process.

STUDENT ENGAGEMENT AND MOTIVATION: SDT

SDT offers a general model of human motivation, with particular attention to external and social factors that can enhance intrinsic motivation as well as specify processes that describe internalization of behaviors that began as externally motivated. The model is well suited to the question of engaging students in course content that is initially of little interest, where student motivation to complete the course is based in external motivation tied to grades and degree requirements but the instructor's goals of engagement and relevance are associated with an internalization of motivation.

In brief, SDT and its associated subtheory, cognitive evaluation theory (CET; Deci & Ryan, 1985), describe the social and environmental factors associated with shifts in intrinsic motivation as well as processes through which externally regulated behaviors can become self-determined and internalized. The model has been applied across a wide range of behavioral domains (see Deci & Ryan, 2000) and is viewed as having particular relevance in educational settings. There are several research examples with relevance to graduate instruction in assessment: Sheldon and Krieger (2007) found that law students who perceived their instructor's style as consistent with SDT principles (discussed later) developed more self-determined motivation; Vansteenkiste, Lens, and Deci (2006) found that when course goals and activities were developed to tap intrinsic motivation, students were more engaged in learning, exhibited stronger persistence, and demonstrated better conceptual thinking. In a prospective study of chemistry students, Black and Deci (2000) found that for students initially low in self-regulated motivation, instruction consistent with SDT principles had a strong relationship to increases in academic performance. Of particular relevance to the domain of student values, Williams and Deci (1996) conducted a longitudinal study examining adoption of biopsychosocial values by medical students. The investigation found that students who perceived their instructors as offering learning conditions consistent with SDT became more self-directed in their learning; reported increases in feelings of competence; and, even at a 2-year follow-up, reported positive and enduring changes in their adoption of psychosocial values relevant to medical practice.

The question of how to engage assessment students in course content, as well as invite them to consider the relevance of testing and assessment to their future work, can be framed within SDT as, "What instructional strategies and what instructor behaviors can facilitate a shift for students from externally regulated motivation to more autonomous motivation?" The core idea underlying SDT is that three innate psychological needs—for autonomy, competence, and relatedness—are important foundations for self-regulated motivation. A succinct summary provided by Ryan and Deci (2000) is informative:

> Not only tangible rewards but also threats, deadlines, directives, pressured evaluations, and imposed goals diminish intrinsic motivation because, like tangible rewards, they conduce toward an external perceived locus of causality. In contrast, choice, acknowledgment of feelings, and opportunities for self-direction were found to enhance intrinsic motivation because they allow people a greater feeling of autonomy (Deci & Ryan, 1985). Field studies have further shown that teachers who are autonomy supportive (in contrast to controlling) catalyze in their students greater intrinsic motivation, curiosity and desire for challenge. . . . Students taught with a more controlling approach not only lose initiative but learn less effectively, especially when learning requires conceptual, creative processing. (pp. 70–71)

The instructor's role is to provide a learning environment that addresses the three core needs, through what SDT terms *autonomy support*, *competence support*, and *relational support*. In an article

that relates SDT to motivation in psychotherapy (Ryan, Lynch, Vansteenkiste, & Deci, 2011), the authors offered definitions and illustrations of these forms of support. Autonomy support, or helping students identify personal goals and reasons for pursuit of learning, is viewed as particularly central to self-regulated motivation. Specific behaviors associated with autonomy support include "(a) offering a meaningful rationale for engaging in the behavior; (b) minimizing external controls such as contingent rewards and punishments; (c) providing opportunities for participation and choice; and (d) acknowledging negative feelings associated with engaging in non-intrinsically motivating tasks" (Ryan et al., 2011, p. 231).

Competence support is closely associated with providing instructional opportunities for students to acquire, practice, and perfect new skills as well as providing assistance and guidance when difficulties are encountered. When students have made a choice to engage, they are more open to skill acquisition. Greater mastery of the material can produce feelings of confidence that can further motivate engagement. This is an area that is likely to be familiar to assessment instructors, as both course content and practicum experience would be closely associated with provision of competence support. The challenge, as noted subsequently, is to provide sufficient coverage of areas where students may feel least competent and that are likely to be areas of knowledge least congruent with their initial attitudes and values. For example, reluctant students may readily integrate the ethical requirements for test use or excel in the interpersonal aspects of test interpretation but may not have integrated the relevance of a predictive validity coefficient or standard error of measurement for a particular testing application.

Relational support refers to creation of a learning environment where students feel respected and valued. This includes an atmosphere of emotional safety in the sense that open discussion of diverse perspectives is permitted, without fear of belittlement or criticism and where there are no "dumb" questions. This is the aspect of SDT instruction that is most relevant to those students who enter the classroom with epistemological and axiological positions that differ from the established traditions of assessment practice. If such students are to engage in assessment course content beyond the external motivation of completing a degree requirement, there needs to be an environment where attitudes and values can be considered openly and respectfully.

Ryan et al.'s (2011) application of SDT to psychotherapy can provide assessment instructors with a more detailed rationale for the points noted earlier as well as catalyze ideas for how SDT can be translated to the classroom. Instructors who choose to implement these principles can be assured that they are in good company: In a 1966 lecture series at Harvard University on "The Psychology of Learning," Carl Rogers called for an approach to teaching that echoes several elements of SDT in asserting that the relationship between teacher and student is critical and that three attitudes characterize effective learning relationships: genuineness, positive regard, and empathy (already familiar as core elements of Rogers's client-centered therapy). In his lecture and consistent with SDT, Rogers argued that teachers who affirm students' self-worth and achieve understanding without judgment will create a classroom atmosphere that encourages self-directed learning. In other words, autonomy support and relational support can set the stage for engagement, which underlies the more familiar movement to competence support.

In the next sections, SDT's tripartite model of autonomy, competence, and relational support guides the description of activities that has been used in this author's own teaching to facilitate student engagement and motivation. Readers are invited to consider the following suggestions as case examples; SDT principles are identified so that instructors can modify the suggested activities to accommodate local interests and priorities.

CLASSROOM ENVIRONMENT AND RELATIONAL SUPPORT

When I serve as instructor of an assessment course, I open with activities designed to provide relational support, through the acknowledgment of potential negative feelings and past anxiety-provoking experiences with testing, and autonomy support, in

communicating that students are not required to agree with the instructor's orientation to testing and that debate on questions of attitudes and values is welcomed. I assert my responsibility to provide an evaluation of students' work and their mastery of course material but indicate that in addition to evaluating their acquired knowledge, I will consider their level of engagement and not their level of agreement. The activities described here can be used during a first or second class meeting and are linked to the assessment competence of self-reflection and self-assessment.

Early Memories

Typically, I begin my first class with an invitation to students to participate in a guided imagery exercise that asks them to recall their first memory of participating in standardized testing: How old were they? What happened? Did they file into a school cafeteria with all their classmates, or visit a "special teacher" in a separate room, where they had to complete a range of tasks and questions? What feelings do they remember having? Curiosity? Anxiety? Boredom? At the time, did they know why they were being asked to complete these tests? Did they ever learn anything about the results? After a few minutes of individual reflection, a classroom discussion is held to explore how early experience can influence attitudes about testing and identify some of the common stereotypic beliefs about testing. Furthermore, students are reminded of what it feels like to be the person being tested, particularly when there is little information about the reasons for testing or the outcome of the assessment. This aspect of the discussion also helps counter the stereotype that client needs are disregarded in testing.

Stereotypes

Either through the first exercise or through a short homework assignment, students are asked to list negative stereotypes about tests that they believe are held by the general public. Targeting stereotypes in the general population creates safety in generating the lists; students feel free to express some of the more outrageous stereotypes without having to claim them as their own views. After the list has been generated, students are asked to select one or two stereotypes that they believe may have some validity or for which they do not have sufficient data/information to challenge the stereotype. One of their assignments is to frame the stereotype in hypothesis-testing terms and, over the course of the term, to collect information germane to assessing their hypothesis. Typically, a small number of stereotypes are identified for testing, which makes it possible for students to work in small groups. In constituting the groups, it is useful to have a range of perspectives that can inform debate. In a closing session, the class reviews and discusses students' emergent views on the stereotypes identified at the outset and often uncovers a change in attitudes. Autonomy support is delivered by giving students permission to express their negative feelings and explore their beliefs by gathering data relevant to the stereotypes. This activity takes minimal class time and can have a significant impact.

OPPORTUNITIES FOR CHOICE AND AUTONOMY SUPPORT

Autonomy support can be delivered by providing student choice for a portion of course assignments. In addition to letting students select some of the measures they will learn in detail, with the direction to choose tests that can be relevant to their future practice, instructors can offer students an opportunity to consider how the therapeutic values of client individuality and culture can be enacted in assessment practice. As one example, a brief written assignment can ask students to link their value for client individuality with a review of psychometric concepts such as standard error of measurement, standard error of estimate, and norms. The relevance of these principles can be discussed for both a dominant-culture client and a marginalized-culture client. Classroom discussions can explore how a Minnesota Multiphasic Personality Inventory—2 profile (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), with representation of scores at 1 or 2 standard deviations above the mean, provides an individualized client description contextualized in a normative group context.

Further examples can include discussion of how the test–retest reliability of the State–Trait Anxiety

Inventory (Spielberger, 1983), or the Beck Depression Inventory (Beck, Steer, & Garbin, 1988) can be particularly informative when conducting assessment with young adults, who are likely to exhibit more volatile emotions. This information is quite basic and is likely covered in many assessment courses; what is different, and delivers autonomy support, is linking the empirical test data to students' concerns with client individuality, thereby countering the stereotype that standardized tests do not treat clients as individuals.

Two other effective examples in this area are the use of career-relevant measures such as the Strong Interest Inventory (Donnay, Morris, Shaubhut, & Thompson, 2004), where client life goals are highly salient, and the use of personality measures such as the Revised NEO Personality Inventory (Costa & McCrae, 1992) with marital conflict, where one can identify personal attributes that both contribute to a couple's relationship as well as cause friction. In both cases, students expand their understanding of how test data can support clients' individualized decision making.

Student motivation, in SDT, is linked to receiving a meaningful rationale for the targeted activity and to minimized external rewards and punishments. Both autonomy and relational support are enhanced by asking students to generate a personal rationale for the achievement of assessment competence and to develop individualized goals for the course. Some aspects of a rationale may still be externally motivated (e.g., the expectations of internship directors in the Association of Psychology Postdoctoral and Internship Centers), whereas others can be intrinsic. For example, students who are less receptive to the use of standardized assessment, but who will be required to use tests during internship or in future employment, may choose a goal of acquiring sufficient knowledge to ensure ethical use of tests and minimization of harm to clients. To increase the salience of such value-congruent goals, instructors can provide in-class examples of how a failure to understand and apply sound measurement principles can place clients at risk.

Finally, the Competencies Work Group description of competence in assessment can provide students with an expanded understanding of the multiple skills and abilities considered important for effective assessment practice. Instructors can identify which elements of competence are prioritized in the current course while inviting students to map the ways they can achieve other areas of competence. Identification of the "O" in KSAO is an effective means of bringing attitudes and values into ongoing conversation.

Competency Support: Integrating Psychometrics and Use of Tests

Competency support consists of providing students with opportunities to acquire and practice skills that build confidence as well as offer targeted assistance when barriers and problems emerge (Ryan et al., 2011). As noted previously, the domain of competency support is aligned with much of the instructional content provided in assessment coursework; however, as the SDT authors noted, development of competency is further enhanced if it is accompanied by autonomy support. Markland, Ryan, Tobin, and Rollnick (2005) found that people were most likely to pursue new competencies when they had already achieved engagement and viewed themselves as acting out of volition, not requirement.

A traditional sequence of assessment coursework moves from measurement and psychometrics to use of tests, to practice experience in practica and internship, but does not always provide opportunities for students to integrate the information presented in each of these settings. A failure of integration underlies the student questions of, "What's a good number for internal consistency reliability?" or "How do I know if what the test manual reports for construct validity fits for my clients?" Some of the key differences between novice and expert thought include the efficient integration of information and a recognition of contradictions; novices tend to memorize facts rather than integrate them and may not recognize contradictions (Adams, Wieman, & Schwartz, 2008).

In my more than 20 years of teaching experience, such failures generally have not been those of cognitive capacity; these questions can come from students who would be assessed as in the top percentiles of academic ability. Instead, it appears that students with minimal interest in measurement

and psychometrics have lacked the motivation to engage with measurement course material and, as a result, remain at a novice level of understanding. These are the students that Helms, Henze, Sass, and Mifsud (2006) described as exhibiting very poor understanding in applying reliability information. In prior coursework, students may memorize course content sufficient to achieve a respectable grade, and can readily list several types of reliability or describe a Spearman–Brown correction, but struggle when asked to apply that knowledge to actual assessment practice. In response to this gap in integration and understanding, there are two types of exercises that can be useful in helping students integrate psychometric information and their use of tests.

Extant research as validity data. As an adjunct to gaining familiarity with a specific measure, students can be asked to identify current research that has used the targeted measure to operationalize variables in a research study (but did not treat the measure as a focus of investigation). The question they are asked to address in a brief paper (e.g. 1–2 pages) is how the identified research can inform their understanding of evidence of validity for the measure's scores with either particular populations or assessment issues. For example, a student might find a research report on career barriers for recent immigrants that used a North American inventory, where scores on the measure successfully predicted which barriers were most closely associated with variables such as social support, language instruction, or past trauma, producing coefficients similar to those obtained in the dominant culture. Students learn that, although caution is still required in using the measure with immigrant populations, existing research can provide evidence for the measure's construct validity.

Classroom exercises linking psychometrics and test use. Exercises that require students to interpret psychometric data in the context of specific testing questions can serve as an assessment of the level of integration students have achieved. Ideally, instruction on specific measures or test practices (e.g., ethics of testing, administration, and interpretation) does not begin until the instructor is confident that students are able to apply and integrate content from their measurement classes; without this step, there is too great a risk that their understanding and integration of further course content will be compromised.

The exercises (see Exhibit 5.1) are designed to be somewhat lighthearted and nonthreatening, consistent with the SDT principles of offering relational support for student engagement (e.g., no dumb questions; you are not being judged on whether you understand this today but will be evaluated on whether you can integrate this information in the future). The exercises can be implemented as individual assignments or in the context of classroom discussion. Topics that are particularly useful, and where examples appear in Exhibit 5.1, are the many uses of correlations and their meaning for specific testing questions, the relevance of test manual information for both test selection and interpretation, and interpretation of various forms of reliability.

The goal for these activities is to create an "aha" moment of recognition; one can almost see the moment when students recognize that the psychometric data provided in test manuals and extant research have relevance for their use of tests with specific clients. Typically, two or three brief exercises are sufficient to produce the desired insight. After all, graduate students are highly competent; once they achieve the necessary integration with respect to either reliability or construct validity, it has been the present author's experience that they readily transfer this integrative knowledge to additional aspects of their measurement and psychometric coursework. Students who would benefit from a review of basic measurement concepts can consult the introductory chapters of the present handbook, including Hubley and Zumbo's overview of psychometrics in assessment (see Volume 1, Chapter 1, this handbook).

Relational Support Redux: Self-Reflection and Competence in Test Interpretation

Professionalism and reflective practice are among the foundational competencies identified by the APA Assessment of Competency Benchmarks Work Group (Fouad et al., 2009) and, considered together, direct test users to take responsibility for competent, ethical practice and to engage in

Exhibit 5.1
Classroom Exercises Linking Psychometrics and Test Use

I. Understanding Correlations in Assessment

For the following exercise, match the circus careers to the measured personality characteristics, based on the psychometric data provided. The correlation coefficient refers to the relationship between the personality factor and the circus career. In the space provided, describe your reasoning for your answers.

Circus Career 1: Lion Tamer
Circus Career 2: Popcorn Vender
Circus Career 3: Human Cannonball
Circus Career 4: Clown

Personality Factor A: Anxiety
Personality Factor B: Flexibility
Personality Factor C: Risk Taking
Personality Factor D: Submissiveness

Circus career	Personality factor	*r*	*p*
1. _____	_____	.29	.05
2. _____	_____	.08	.05
3. _____	_____	.54	.09
4. _____	_____	–.41	.01

Answers: 1. Clown, Flexibility; 2. Popcorn Vender, Anxiety; 3. Human Cannonball, Risk-Taking; 4. Lion Tamer, Submissiveness

II. Psychometrics and Test Selection

A national department store chain has asked you to evaluate the effectiveness of its 10-week relaxation/anger management program for employees who work in customer service and returns. You find a 20-item measure called the Irritability Index that you hope will identify which employees do and do not benefit from the intervention. You track down several research articles that provide the following information on the psychometric properties of the scale: What does this information tell you about the usefulness of the Irritability Index for evaluating the success of the intervention program? Please organize your answers by letter (e.g., "a," "b,").

a. A well-established Hostility Scale, with 125 items, has a correlation of $r = .79$ with the Irritability Index.
b. Correlations between the Irritability Index and clinician ratings of client well-being range from –.31 to –.36.
c. The Irritability Index test manual reports a 3-week test–retest reliability as $r = .65$, compared with $r = .81$ for the Hostility Scale.
d. The Irritability Index was developed on a large Australian sample of male high school students who had been referred to school principals for disciplinary action.
e. When the Irritability Index was administered to players in the National Football League, the correlation between scores and number of penalties received per game was $r = .39$.
f. The Irritability Index is available in a Spanish-language translation; the correlation between the original measure and the translation, when administered to fully bilingual test takers, was $r = .51$.

III. Reliability

(A) One measure of social support, Test Q, considers emotional, financial, and practical forms of assistance and support. Another measure, Test R, was designed to test only emotional support. Both are well constructed, and each has evidence for validity for its scores. Which reliability coefficient is a likely candidate to describe the internal consistency of Test Q and Test R? Please explain your reasoning.

Test Q:	.27	.65	.89
Test R:	.27	.65	.89

Answers: Test Q = .65, as there is still an underlying construct of social support; Test R = .89, given the limited focus on one form of social support.

(B) You have a well-respected measure of *state* or situational anxiety. If you collect data and calculate test–retest reliability coefficients for 3 days, 3 weeks, and 3 months, what pattern would you expect to see in the correlations? For a measure of *trait* anxiety? Please explain your reasoning.

	3 day	3 week	3 month
a.	.74	.46	.37
b.	.87	.91	.52
c.	.87	.86	.75
d.	.64	.67	.66

Answers: State anxiety = a; the coefficient for 3-day test–retest is acceptable and, as expected, the pattern over time is less stable; Trait anxiety = c; the values of the coefficients demonstrate stability yet show an expected decrease over time. Option b: the pattern reflects an implausible increase in test–retest at 3 weeks; Option d: the pattern reflects an implausible lack of change over time.

(C) Adolescents are known for having volatile emotions. Which pair of correlations for adolescents and adults do you think is most likely to describe the test–retest coefficients for scores on a well-known, well-established depression inventory? Please explain your reasoning.

	Adolescents	Adults
a.	.61	.74
b.	.92	.95
c.	.76	.60
d.	.20	.27

Answers: The coefficient for adolescents is expected to be lower, reflecting volatility; among the three options that meet that requirement, the correct answer is a, as test–retest coefficients in either the .20s or the .90s are implausible for a depression measure.

personal and professional self-assessment with respect to their practice. For students, this aspect of competency development requires that they understand and accept their ethical and professional role in the use of tests. This is an area where relational support is particularly relevant for the development of student competence. Acquiring the professional mien of a psychologist with the associated skills is a developmental challenge because it takes time to develop the confidence to negotiate difficult encounters; this is as true of assessment as it is of psychotherapy. Students whose philosophy of science and worldview are aligned with postmodern and humanistic perspectives may express opposition to the information that tests provide about client areas of difficulty, viewing this as pathologizing, while voicing a desire to emphasize client strengths. In response, an instructor can point out that, in psychotherapy and assessment, no one benefits from a one-sided portrait and that, in fact, avoidance of difficult material is not consistent with the principles of respect and open communication.

Fowler (1998) has pointed out that for students, learning to conduct assessment has much in common with learning to be a psychotherapist and that, in both cases, student anxieties and role uncertainty can create barriers to effective learning. Although Fowler has framed this in the language of resistance and projection, his comments also are relevant for students whose philosophy of science and worldview may be at odds with traditional assessment practice. Specifically, he has identified four areas that students experience as difficulties: "the wish and fear of the expert voice; balancing morbid vs. Pollyannaish interpretations; using oneself as an instrument of assessment; and integration and communication" (p. 34).

These issues might be most effectively framed as ethical responsibilities, which is also consonant with the values/axiology of students who identify with constructivist and postmodern perspectives. A psychology trainee who wants to honor multiple perspectives on reality can find it hard to defend an exclusive focus on positive attributes and client strengths; not disclosing information on areas of concern can be viewed as being as unethical and disrespectful as an exclusive focus on difficulties and pathology. At the same time, it is important to acknowledge the anxiety and uncertainty that students experience at the prospect of communicating difficult or unwelcome test results. Fowler (1998) noted that students must be assisted to avoid "Barnum" interpretations as they find a balance in describing test takers' strengths and weaknesses.

Many students who find this challenging also have received a thorough grounding in psychotherapeutic skills but may not have made the connection in understanding how those skills are relevant in assessment practice. To assist with that integration, instructors might want to use one or more role-play scenarios that bring these challenges into sharp focus and provide students with an opportunity to practice and develop a sense of competence. The following examples have all elicited student discomfort

over the prospect of communicating tests results potentially unwelcome to clients: communicating cognitive ability test results to a career client whose abilities are modest relative to career goals; communicating the diagnostic results of a child's learning disability assessment to parents who have high expectations for their child's school achievements; communicating personality test results to clients concerned about their mental health status.

In each case, the class works from mock test data, with sufficient information to make the case realistic. Before initiating the role play, students are asked to consider the following questions: What is the goal of testing in this case? What questions or anxieties are likely to be in the client's mind? What are the short-term and long-term risks of over- or underemphasizing particular aspects of the results? What values and stance toward clients are conveyed in choosing to provide a comprehensive versus "edited" version of the relevant test results?

The role-play exercise provides autonomy support by giving students an opportunity to voice their discomfort and worries about competence and by inviting them to define what constitutes an effective, respectful test interpretation. Relational support is enacted by labeling the task as challenging and by assuring students that awkwardness and mistakes are to be expected. Competency support is delivered by helping the group to generate concrete language and phrases that can be used to communicate difficult results and then providing opportunities to practice to reduce discomfort and increase confidence.

The Fourth Quadrant: Collaborative and Therapeutic Assessment

Assessment instructors who want to engage students with humanistic and postmodern values and attitudes are strongly encouraged to incorporate a unit on collaborative and therapeutic models of assessment (Duckworth, 1990; Finn & Tonsager, 1997; Fischer, 2000; see also Chapter 28, this volume). Emerging in the 1990s, a range of approaches to assessment practice have emphasized collaboration with test takers on everything from the decision of whether to use standardized tests for information gathering to active engagement in interpretation and application of results. Furthermore, therapeutic assessment models view testing as an active intervention in its own right rather than as simply an aid to information gathering. Evidence in support of this claim is substantive; a meta-analysis (Poston & Hanson, 2010) found that collaborative assessment procedures have a demonstrated, positive effect on client treatment, with an overall effect size of $d = 0.423$. The scope of this chapter does not permit a detailed discussion of the models themselves; interested readers are encouraged to consult Chapter 26 of this volume as well as Finn (2007) and Fischer (1994) for detailed examples.

The processes and values of collaborative assessment are highly consistent with the worldview espoused by what have been described as "reluctant students," those who want to prioritize collaborative relationships with test takers, place client needs at the center of testing, and argue for the importance of highly individualized, contextualized information. These factors help define the lower righthand corner of the epistemology/axiology grid described earlier; introducing students to models of therapeutic assessment provides an example of how they can conduct assessment in a manner congruent with their professional worldview. Although differing in some areas of emphasis, Finn and Tonsager (1997) described models of therapeutic assessment as sharing the elements of "(a) developing and maintaining empathic connections with clients, (b) working collaboratively with clients to develop individualized assessment goals, and (c) sharing assessment results with clients" (p. 278). These points echo a perspective advanced in 1990 by Duckworth, who described a counseling psychology perspective on testing as being for the benefit of the test taker, not just the psychologist, and that the goal of testing is the empowerment of the individual through provision of increased self-knowledge and skills.

When introducing collaborative and therapeutic assessment practice to students, it is important to point out that, with respect to the grid in Figure 5.1, these models reflect an integration of empirical and individualized perspectives rather than being fully positioned at one pole of the *y*-axis. The critical point within this observation is that, to engage one's clients in a genuinely therapeutic and collaborative

assessment process, one has to begin with valid and reliable assessment information, associated with the empirical end of the continuum. In the models noted, standardized test data are interpreted in the context of highly individualized client goals and circumstances; and, the most effective and meaningful test interpretations will be provided by psychologists who have a knowledgeable, sophisticated understanding of the psychometric basis of a measure. In other words, the procedures of therapeutic assessment can provide a meaningful rationale for achieving competency in the psychometric building blocks that underlie these approaches. There is now good evidence that clients benefit from a collaborative assessment experience when accompanied by highly individualized, involving feedback (Poston & Hanson, 2010). To achieve this therapeutic goal, it is still necessary to begin with a solid understanding of the measures and the trustworthiness of the information they can offer to clients. However, the promise of therapeutic impact can provide students with a meaningful rationale for the acquisition of psychometric and measurement knowledge.

CONCLUSION

The field has seen a renewal of interest in assessment instruction, spurred in part by decreases in the use of standardized testing by practicing psychologists and concerns over reductions in curricular attention to tests and measurement. The argument in this chapter is that some student ambivalence about assessment may be attributable to differences with instructors on a basic epistemological and axiological worldview and describes ways to engage "reluctant students" in assessment courses.

It is worth noting that student disinterest in assessment has been a perennial source of discussion, even angst. In 1972, Leo Goldman remarked that students "typically have objected to what they regard as excessive time and attention given to tests and measurement courses" (p. 213) and in 1994, Dale Prediger responded, "Amen! And it will ever be so" (p. 228). Psychology's greater awareness of diverse philosophy of science paradigms may provide the means to disprove Prediger's pessimistic prediction by providing ways to help students bridge the analytic and intuitive worldviews. To build that bridge, instructors need to communicate respect for student worldviews and provide alternative models of assessment congruent with students' therapeutic values. A key benefit of this approach is that "reluctant students" are given an opportunity to learn for themselves that achieving competence in psychometrics, measurement, and test use can expand their therapeutic effectiveness. With experience, students learn that assessment competence is a means for providing trustworthy information that can enhance their clients' self-knowledge and choices.

References

Adams, W., Wieman, C., & Schwartz, D. (2008). *Teaching expert thinking*. Retrieved from http://www.cwsei.ubc.ca/resources/files/Teaching_Expert_Thinking.pdf

Alfonso, V. C., & Pratt, S. I. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 326–344). New York, NY: Guilford Press.

American Psychological Association. (2006). *Report of the APA Task Force on the Assessment of Competence in Professional Psychology*. Washington, DC: Author.

Aschieri, F., Finn, S. E., & Bevilacqua, P. (2010). Therapeutic assessment and epistemological triangulation. In V. Cigoli & M. Gennari (Eds.), *Close relationships and community psychology: An international perspective* (pp. 241–253). Milan, Italy: Franco Angeli.

Babarik, P. (1980). "What do they really want?" A survey of undergraduates' preferences and aversions in psychology. *Canadian Psychology/Psychologie canadienne, 21*, 84–86. doi:10.1037/h0081097

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years later. *Clinical Psychology Review, 8*, 77–100. doi:10.1016/0272-7358(88)90050-5

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57*, 717–726. doi:10.1002/jclp.1044

Black, A. E., & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education, 84*, 740–756. doi:10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory (MMPI–2). Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Cacioppo, J. T. (2004). Common sense, intuition and theory in personality and social psychology. *Personality and Social Psychology Review, 8*, 114–122. doi:10.1207/s15327957pspr0802_4

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi:10.1037/0735-7028.31.2.141

Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment, 76*, 18–47. doi:10.1207/S15327752JPA7601_2

Corey, G. (2009). *Theory and practice of counseling and psychotherapy* (8th ed.). Belmont, CA: Thomsen Brooks/Cole.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Curry, K. T., & Hanson, W. E. (2010). National survey of psychologists' test feedback training, supervision and practice: A mixed methods study. *Journal of Personality Assessment, 92*, 327–336. doi:10.1080/00223891.2010.482006

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York, NY: Plenum Press.

Deci, E. L., & Ryan, R. M. (2000). The "what" and the "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268. doi:10.1207/S15327965PLI1104_01

Donnay, D. A. C., Morris, M. L., Shaubhut, N. A., & Thompson, R. C. (2004). *Strong Interest Inventory manual: Research, development and strategies for interpretation.* Mountain View, CA: Consulting Psychologists Press.

Duckworth, J. (1990). The counseling approach to the use of testing. *The Counseling Psychologist, 18*, 198–204. doi:10.1177/0011000090182002

Eisman, E., Dies, R., Finn, S. E., Eyde, L., Kay, G. G., Kubiszyn, T., . . . Moreland, K. (1998). *Problems and limitations in the use of psychological assessment in contemporary healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part II.* Washington, DC: American Psychological Association.

Finn, S. E. (2007). *In our clients' shoes: Theory and techniques of therapeutic assessment.* Mahwah, NJ: Erlbaum.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment, 4*, 278–287. doi:10.1037/1040-3590.4.3.278

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9*, 374–385. doi:10.1037/1040-3590.9.4.374

Fischer, C. T. (1994). *Individualizing psychological assessment.* Mahwah, NJ: Erlbaum.

Fischer, C. T. (2000). Collaborative, individualized assessment. *Journal of Personality Assessment, 74*, 2–14. doi:10.1207/S15327752JPA740102

Fouad, N. A., Grus, C. L., Hatcher, R. L., Kaslow, N. J., Smith Hutchings, P., Madson, M. B., . . . Crossman, R. E. (2009). Competency benchmarks: A model for understanding and measuring competence in professional psychology across training levels. *Training and Education in Professional Psychology*, *3*(4, Suppl.), S5–S26. doi:10.1037/a0015832

Fowler, J. C. (1998). The trouble with learning personality assessment. In L. Handler & M. J. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 31–41). Mahwah, NJ: Erlbaum.

Goldman, L. (1972). Tests and counseling: The marriage that failed. *Measurement and Evaluation in Guidance, 4*, 213–220.

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Testing Cronbach's alpha reliability coefficients as data in counseling research. *The Counseling Psychologist, 34*, 630–660. doi:10.1177/0011000006288308

Heppner, P. P., & Claiborn, C. D. (1989). Social influence research in counseling: A review and critique. *Journal of Counseling Psychology, 36*, 365–387. doi:10.1037/0022-0167.36.3.365

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, NY: Oxford University Press.

Kaslow, N. J. (2004). Competencies in professional psychology. *American Psychologist, 59*, 774–781. doi:10.1037/0003-066X.59.8.774

Kaslow, N. J., Borden, K. A., Collins, F. L., Jr., Forrest, L., Illfelder-Kaye, J., Nelson, P. D., . . . Willmuth, M. E. (2004). Competencies conference: Future directions in education and credentialing in professional psychology. *Journal of Clinical Psychology*, *60*, 669–712. doi:10.1002/jclp.20016

Kaslow, N. J., Grus, C. L., Campbell, L. F., Fouad, N. A., Hatcher, R. L., & Rodolfa, E. R. (2009). Competency assessment toolkit for professional psychology. *Training and Education in Professional Psychology*, *3*(4, Suppl.), S27–S45. doi:10.1037/a0015833

Koocher, G. S., & Keith-Spiegel, P. (2008). *Ethics in psychology and the mental health professions: Standards and cases* (3rd ed.). New York, NY: Oxford University Press.

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Milville, M. L., Kerns, R., . . . Benton, S. A. (2004). Achieving competence in psychological assessment: Directions for education and training. *Journal of Clinical Psychology*, *60*, 725–739. doi:10.1002/jclp.20010

Levin, B. (2008). *Thinking about knowledge mobilization.* Vancouver, British Columbia, Canada: Canadian Council on Learning and the Social Sciences and Humanities Research Council of Canada. Retrieved from http://www.oise.utoronto.ca/Conferences_Presentations_Publications/index.html

LoBello, S. G., & Zachar, P. (2007). Psychological test sales and internet auctions: Ethical considerations for dealing with obsolete or unwanted test materials. *Professional Psychology: Research and Practice, 38*, 68–70. doi:10.1037/0735-7028.38.1.68

Markland, D., Ryan, R. M., Tobin, V. J., & Rollnick, S. (2005). Motivational interviewing and self-determination theory. *Journal of Social and Clinical Psychology, 24*, 811–831. doi:10.1521/jscp.2005.24.6.811

Mintz, L. B., Jackson, A. P., Neville, H. A., Illfelder-Kaye, J., Winterowd, C. L., & Loewy, M. I. (2009). The need for a counseling psychology model training values statement addressing diversity. *The Counseling Psychologist, 37*, 644–675. doi:10.1177/0011000009331931

Neimeyer, R. A. (1995). An invitation to constructivist psychotherapies. In R. A. Neimeyer & M. J. Mahoney (Eds.), *Constructivism in psychotherapy* (pp. 1–8). Washington, DC: American Psychological Association. doi:10.1037/10170-018

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment, 22*, 203–212. doi:10.1037/a0018679

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations. *Journal of Vocational Behavior, 21*, 259–287. doi:10.1016/0001-8791(82)90036-7

Prediger, D. J. (1994). Tests and counseling: The marriage that prevailed. *Measurement and Evaluation in Counseling and Development, 26*, 227–234.

Prediger, D. J., & Swaney, K. B. (2004). Work task dimensions underlying the world of work: Research results for diverse occupational databases. *Journal of Career Assessment, 12*, 440–459. doi:10.1177/1069072704267737

Rodolfa, E. R., Bent, R. J., Eisman, E., Nelson, P. D., Rehm, L., & Ritchie, P. (2005). A Cube model for competency development: Implications for psychology educators and regulators. *Professional Psychology: Research and Practice, 36*, 347–354. doi:10.1037/0735-7028.36.4.347

Rogers, C. R. (1966, April 13). The psychology of learning. *Harvard Crimson*. Retrieved from http://www.thecrimson.com/article/1966/4/13/carl-r-rogers-spells-out-new

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78. doi:10.1037/0003-066X.55.1.68

Ryan, R. M., Kuhl, J., & Deci, E. L. (1997). Nature and autonomy: Organizational view of social and neurobiological aspects of self-regulation in behavior and development. *Development and Psychopathology, 9*, 701–728. doi:10.1017/S0954579497001405

Ryan, R. M., Lynch, M. F., Vansteenkiste, M., & Deci, E. L. (2011). Motivation and autonomy in counseling, psychotherapy and behavior change: A look at theory and practice. *The Counseling Psychologist, 39*, 193–260. doi:10.1177/0011000009359313

Sheldon, K. M., & Krieger, L. S. (2007). Education on law students: A longitudinal test of self-determination theory. *Personality and Social Psychology Bulletin*, *33*, 883–897. doi:10.1177/0146167207301014

Sherif, C. W., Sherif, M. S., & Nebergall, R. E. (1965). *Attitude and attitude change.* Philadelphia, PA: W. B. Saunders.

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change.* New Haven, CT: Yale University Press.

Spielberger, C. D. (1983). *Manual for the State–Trait Anxiety Inventory: STAI (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psycholol gical assessment training in graduate and professional schools. *Journal of Personality Assessment, 77*, 398–407. doi:10.1207/S15327752JPA7703_02

Strong, S. R. (1968). Counseling: An interpersonal influence process. *Journal of Counseling Psychology, 15*, 215–224. doi:10.1037/h0020229

Tinsley, H. E. A., & Bradley, R. W. (1986). Test interpretation. *Journal of Counseling and Development, 64*, 462–466. doi:10.1002/j.1556-6676.1986.tb01166.x

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist, 41*, 19–31. doi:10.1207/s15326985ep4101_4

Watkins, C. E., & Campbell, V. L. (1990). Testing and assessment in counseling psychology: Contemporary developments and issues. *The Counseling Psychologist, 18*, 189–197. doi:10.1177/0011000090182001

Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology, 70*, 767–779. doi:10.1037/0022-3514.70.4.767

CHAPTER 6

LEGAL ISSUES IN CLINICAL AND COUNSELING TESTING AND ASSESSMENT

Elizabeth V. Swenson

The law that affects psychological testing comes from statutes, the holdings of legal cases, and the rules emanating from administrative agencies such as the Equal Employment Opportunities Commission. These represent the three branches of government. They reside in both the federal and the state government. The focus of this chapter is federal law, instead of the many individual states' laws. In addition, much that relates to the law comes from professional ethics. Although the states have laws that cover psychologists, the ethical issues discussed are those of the American Psychological Association (APA) Ethical Principles of Psychologists and Code of Conduct (APA, 2010), again, because of their national applicability and because they form the basis for many state standards and statutes.

When thinking of legal issues in psychological testing, one is immediately drawn to the famous Supreme Court cases in employment testing. A well-known example is *Griggs v. Duke Power* (1971), where the Court held that psychological tests that have a disparate impact on a minority group must be closely related to the job for which they select (see Volume 1, Chapter 38, this handbook). Educational tests, as well, have had their fair share of landmark legal decisions, such as *Larry P. v. Riles* (1984), where the Ninth Circuit Court of Appeals held that IQ tests were not free of cultural and racial bias and therefore could not be the sole selection procedure for special education classes (see Volume 3, Chapter 12, this handbook). Legal cases are not as well known when one considers psychological diagnostic and personality testing and assessment.

FEDERAL LAWS AFFECTING PSYCHOLOGICAL TESTING

Title VII of the 1964 Civil Rights Act has had a far-reaching effect on psychological testing. Although designed to prohibit discrimination in employment, the act has had the more general effect of focusing psychologists and test makers on the validity and use of psychological tests with protected groups. The 1978 Uniform Guidelines on Employee Selection Procedures have had a trickle-down effect on practitioners who use psychological tests for other purposes, emphasizing fairness and the lack of test bias. The Equal Employment Opportunity Commission (EEOC) Uniform Guidelines emphasize that employment testing needs to be closely related to the criteria for selection.

The Health Insurance Portability and Accountability Act (HIPAA) was enacted in 1996 to protect the privacy of an individual's personal health information and data. A person's Protected Health Information (PHI) includes personally identifiable health records that an individual has the right to inspect, copy, and have transferred to a person of their choice. This description applies to psychological test data.

The Family Educational Rights and Privacy Act of 1974 (FERPA) gives the right of privacy, inspection, and review of student records to individuals over the age of 18 and to parents of children under 18. Included in student records are psychological test data that are possessed by an educational institution. FERPA allows many exceptions.

DOI: 10.1037/14048-006
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

THE APA ETHICS CODE

As with most activities in psychology, testing and assessment for clinical and counseling information lends itself to a variety of legal and ethical issues. At the most fundamental level, legal issues are anchored in the ethics of the profession of psychology.

To do the best assessment, and to give the best service to the community and to the profession, investigators need to behave ethically. Much has been written about the ethics of assessment in the behavioral sciences. For assessment in clinical and counseling psychology specifically, the guidelines are set forth clearly in Section 9, Assessment, of the APA Ethical Principles of Psychologists and Code of Conduct (referred to hereinafter as Ethics Code; APA, 2010). This document is in its 10th version and was designed specifically to meet "ethical challenges in the new millennium" (Fisher, 2003, p. xxv). Although it can be supplemented by other documents, the Ethics Code is the foundation for this chapter on legal issues in psychological assessment.

The Ethics Code consists of 10 sections. Within Standard 9, there are 11 substandards covering various aspects of assessment, although other standards in the Code apply to psychological assessment as well. The standards are preceded by five General Ethical Principles that underlie the standards in the Ethics Code and are aspirational in nature. These are: Beneficence and Nonmaleficence, Fidelity and Responsibility, Integrity, Justice, and Respect for People's Rights and Dignity (Exhibit 6.1). These principles are considered to be the moral basis for the Ethics Code and are similar to those well known in bioethics (Beauchamp & Childress, 2001). In the application of the standards of the Ethics Code to specific situations in assessment, often the correct answer is not readily apparent. The psychologist/decision maker then needs to apply the General Ethical Principles to aid in decision making, after which she or he should consult a colleague and document the process. Although the General Principles are aspirational in nature, the standards from the Ethics Code are enforceable. Exhibit 6.2 presents the entire Standard 9 of the Ethics Code.

For legal purposes, the Ethics Code is important because it is the single most important statement of

Exhibit 6.1
The General Principles From the Ethics Code that Guide and Inspire Psychologists in All Their Work

Principle A: Beneficence and Nonmaleficence
Psychologists strive to benefit those with whom they work and take care to do no harm. In their professional actions, psychologists seek to safeguard the welfare and rights of those with whom they interact professionally.. . . Because psychologists' scientific and professional judgments and actions may affect the lives of others, they are alert to and guard against personal, financial, social, organizational, or political factors that might lead to misuse of their influence. Psychologists strive to be aware of the possible effect of their own physical and mental health on their ability to help those with whom they work.

Principle B: Fidelity and Responsibility
Psychologists establish relationships of trust with those with whom the work. They are aware of their professional and scientific responsibilities to society and to the specific communities in which they work. Psychologists uphold professional standards of conduct.

Principle C: Integrity
Psychologists seek to promote accuracy, honesty, and truthfulness in the science, teaching, and practice of psychology. In these activities psychologists do not steal, cheat, or engage in fraud, subterfuge, or intentional misrepresentation of fact.

Principle D: Justice
Psychologists recognize that fairness and justice entitle all persons to access to and benefit from the contributions of psychology. . . . Psychologists exercise reasonable judgment and take precautions to ensure that their potential biases . . . and the limitations of their expertise do not lead or condone unjust practices.

Principle E: Respect for People's Rights and Dignity
Psychologists respect the dignity and worth of all people, and the rights of individuals to privacy, confidentiality, and self-determination. Psychologists are aware that special safeguards may be necessary to protect the rights and welfare of persons or communities whose vulnerabilities impair autonomous decision making. Psychologists are aware of and respect cultural, individual, and role differences, including those based on age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, and socioeconomic status.

Note. From *Ethical Principles of Psychologists and Code of Conduct* (2010), by the American Psychological Association, Washington, DC: Author. Copyright 2010 by the American Psychological Association.

Exhibit 6.2
Standard 9: Assessment

9.01 Bases for Assessments

(a) Psychologists base the opinions contained in their recommendations, reports and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings. (See also Standard 2.04, Bases for Scientific and Professional Judgments.)

(b) Except as noted in 9.01c, psychologists provide opinions of the psychological characteristics of individuals only after they have conducted an examination of the individuals adequate to support their statements or conclusions. When, despite reasonable efforts, such an examination is not practical, psychologists document the efforts they made and the result of those efforts, clarify the probable impact of their limited information on the reliability and validity of their opinions, and appropriately limit the nature and extent of their conclusions or recommendations. (See also Standards 2.01, Boundaries of Competence, and 9.06, Interpreting Assessment Results.)

(c) When psychologists conduct a record review or provide consultation or supervision and an individual examination is not warranted or necessary for the opinion, psychologists explain this and the sources of information on which they based their conclusions and recommendations.

9.02 Use of Assessments

(a) Psychologists administer, adapt, score, interpret or use assessment techniques, interviews, tests or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques.

(b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation.

(c) Psychologists use assessment methods that are appropriate to an individual's language preference and competence, unless the use of an alternative language is relevant to the assessment issues.

9.03 Informed Consent in Assessments

(a) Psychologists obtain informed consent for assessments, evaluations, or diagnostic services, as described in Standard 3.10, Informed Consent, except when (1) testing is mandated by law or governmental regulations; (2) informed consent is implied because testing is conducted as a routine educational, institutional or organizational activity (e.g., when participants voluntarily agree to assessment when applying for a job); or (3) one purpose of the testing is to evaluate decisional capacity. Informed consent includes an explanation of the nature and purpose of the assessment, fees, involvement of third parties, and limits of confidentiality and sufficient opportunity for the client/patient to ask questions and receive answers.

(b) Psychologists inform persons with questionable capacity to consent or for whom testing is mandated by law or governmental regulations about the nature and purpose of the proposed assessment services, using language that is reasonably understandable to the person being assessed.

(c) Psychologists using the services of an interpreter obtain informed consent from the client/patient to use that interpreter, ensure that confidentiality of test results and test security are maintained, and include in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, discussion of any limitations on the data obtained. (See also Standards 2.05, Delegation of Work to Others; 4.01, Maintaining Confidentiality; 9.01, Bases for Assessments; 9.06, Interpreting Assessment Results; and 9.07, Assessment by Unqualified Persons.)

9.04 Release of Test Data

(a) The term test data refers to raw and scaled scores, client/patient responses to test questions or stimuli, and psychologists' notes and recordings concerning client/patient statements and behavior during an examination. Those portions of test materials that include client/patient responses are included in the definition of test data. Pursuant to a client/patient release, psychologists provide test data to the client/patient or other persons identified in the release. Psychologists may refrain from releasing test data to protect a client/patient or others from substantial harm or misuse or misrepresentation of the data or the test, recognizing that in many instances release of confidential information under these circumstances is regulated by law. (See also Standard 9.11, Maintaining Test Security.)

(b) In the absence of a client/patient release, psychologists provide test data only as required by law or court order.

9.05 Test Construction

Psychologists who develop tests and other assessment techniques use appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, reduction or elimination of bias, and recommendations for use.

(*Continued*)

Exhibit 6.2
Continued

9.06 Interpreting Assessment Results

When interpreting assessment results, including automated interpretations, psychologists take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists' judgments or reduce the accuracy of their interpretations. They indicate any significant limitations of their interpretations. (See also Standards 2.01b and c, Boundaries of Competence, and 3.01, Unfair Discrimination.)

9.07 Assessment by Unqualified Persons

Psychologists do not promote the use of psychological assessment techniques by unqualified persons, except when such use is conducted for training purposes with appropriate supervision. (See also Standard 2.05, Delegation of Work to Others.)

9.08 Obsolete Tests and Outdated Test Results

(a) Psychologists do not base their assessment or intervention decisions or recommendations on data or test results that are outdated for the current purpose.

(b) Psychologists do not base such decisions or recommendations on tests and measures that are obsolete and not useful for the current purpose.

9.09 Test Scoring and Interpretation Services

(a) Psychologists who offer assessment or scoring services to other professionals accurately describe the purpose, norms, validity, reliability, and applications of the procedures and any special qualifications applicable to their use.

(b) Psychologists select scoring and interpretation services (including automated services) on the basis of evidence of the validity of the program and procedures as well as on other appropriate considerations. (See also Standards 2.01b and c, Boundaries of Competence.)

(c) Psychologists retain responsibility for the appropriate application, interpretation, and use of assessment instruments, whether they score and interpret such tests themselves or use automated or other services.

9.10 Explaining Assessment Results

Regardless of whether the scoring and interpretation are done by psychologists, by employees or assistants, or by automated or other outside services, psychologists take reasonable steps to ensure that explanations of results are given to the individual or designated representative unless the nature of the relationship precludes provision of an explanation of results (such as in some organizational consulting, preemployment or security screenings, and forensic evaluations), and this fact has been clearly explained to the person being assessed in advance.

9.11 Maintaining Test Security

The term test materials refers to manuals, instruments, protocols, and test questions or stimuli and does not include test data as defined in Standard 9.04, Release of Test Data. Psychologists make reasonable efforts to maintain the integrity and security of test materials and other assessment techniques consistent with law and contractual obligations, and in a manner that permits adherence to this Ethics Code.

Note. From *Ethical Principles of Psychologists and Code of Conduct* (2010), by the American Psychological Association, Washington, DC: Author. Copyright 2010 by the American Psychological Association.

the standard of care for psychologists. As refined further by the states' individual psychology laws, violation of the standard of care has two legal ramifications. The first is that it can result in sanctions by the state licensing board. These sanctions can range from a warning or reprimand to loss of one's license to practice psychology. The second legal ramification is that breach of the standard of care can result in malpractice. Also worth noting is that peer review groups, such as the APA Ethics Committee, can impose sanctions such as expulsion from the APA or required practice monitoring.

Malpractice

For malpractice to occur, four elements must be present. First, there must be a duty of care. A duty arises when a psychologist takes on a client or patient. The legal duty is from the psychologist to the client. Second, there must be a breach of the duty or a lapse from the standard of care. Third,

there needs to be an injury to the client. Finally, the breach of the duty must be the proximate cause of the injury to the client. In other words, the injury must have been caused by the psychologist's violation of the Ethics Code or whatever other standard has been adopted by the jurisdiction in question. According to Gutheil (2009), the leading causes of action for malpractice for psychiatrists that might have an assessment issue include suicide, failure to maintain appropriate confidentiality, and boundary issues. Many of these cases are accompanied by some sort of incompetence as well. In the following excerpts, case examples are hypothetical except for those that are included in the discussion of an actual legal case.

Two hypothetical, but not unrealistic, examples follow.

> Margo Mayem, PsyD, teaches a course in psychological assessment to a class of advanced undergraduate psychology majors. She feels quite strongly that the only way students can really understand psychological testing is to actually administer selected tests to themselves, score them, and discuss the results. She has done this with several tests such as the Myers–Briggs Type Indicator (MBTI) and the Strong Interest Inventory (SII). The students, however, were not as interested in these tests as she had hoped. To enliven the class a bit, Dr. Mayem decided to give each student the Minnesota Multiphasic Personality Inventory—2 (MMPI–2). Because of severe time constraints, the students were to take the test home with them and then to fill out the answer sheet. Although they were told to read the directions carefully and then to settle in a quiet room and take the test in one sitting, it was clear that some students did not do this. In fact, it was reported in the next class that some of the "sillier" test items were texted back and forth among the students and their friends. Most students turned in their answer sheets and test booklets at the next class. Dr. Mayem, realizing that the administration of the MMPI had gotten out of hand, decided to collect the answer sheets and to score them herself. To her horror, two of the 14 students exhibited severe psychopathology. One woman had elevated 4 (psychopathic deviate) and 8 (schizophrenia) Scales, and a young man had a severely elevated 2 (depression) and slightly elevated 6 (paranoia) Scale. Dr. Mayem quickly realized that she was in over her head and that she had no need or desire to know so much about her students. Wondering how to handle the class at this point, she decided to consult her department chair. His advice was to discuss the test results generally but not to return results to individual students. The class raised an uproar when told they would not learn their results. Dr. Mayem blurted out to the class, "Two of you scored alarmingly. I can't reveal this to you." As it turned out, all of the students felt that Dr. Mayem was pointing her finger at them. To some extent, all the students felt uneasy about their relationship to their professor, and four students sought counseling for what they thought was severe psychopathology.

Dr. Mayem's behavior is problematic under several sections of Standard 9 of the Ethics Code. First, Standard 9.02(a) states that tests are to be administered and used in a manner and for purposes that are appropriate. Administering this test to a whole class of undergraduates is questionable, as is letting them take the test home. Considering how damaging the results could be to individual students, it is also clear that Standard 9.03(a), Informed Consent in Assessments, was not applied. Students should have been told, under Standard 9.06, that it might be impossible to make a valid interpretation of the results because of the random method of test administration. Under Standard 9.10, students had the right to a reasonable attempt to explain the results. This needed to be done individually, not in the

class, with a huge disclaimer about the poor test-taking conditions. Students believed in advance that they would learn the results, and Dr. Mayem was responsible for communicating their test results to them (see also Chapter 3, this volume). Finally Standard 9.11, Maintaining Test Security, was violated. Of these standards, the failure to receive informed consent along with the failure to give results, qualified by the lack of a standardized test-taking condition, are the most likely to have psychologically injured the students. In fact, one of the students did sue the university and the professor for damages, and another lodged a complaint with the licensing board. The teacher had a duty to the students that she breached by violating ethics standards regarding assessment. As a result, some students were injured. All the elements of malpractice are present.

> Martin Rednik, PhD, read about a weekend workshop designed to train mental health professionals on how to recognize victims of torture and to testify at their asylum hearings. He attended the workshop and developed an interest in expanding his practice to include assessing and giving career planning advice to immigrants. (After learning about torture, he quickly decided that dealing with these tortured people was not for him.) Recalling some assessment procedures he had learned about in graduate school some 20 years ago, he put together a new practice area that would be both interesting and untouched by managed care. He marketed his practice by sending flyers to immigration attorneys and immigrant neighborhood outreach programs as well as advertising in foreign language newspapers and newsletters. Dr. Rednik's protocol consisted of giving each person the Wechsler Adult Intelligence Scale—III (WAIS–III) followed by the Rorschach Inkblot Test and the Strong Interest Inventory (SII). Knowing that a person should be tested in their preferred language, Dr. Rednik hired translators in Hmong, Farsi, and Swahili. Testing with the translators went well, but the results were amazing to Dr. Rednik. Wide discrepancies existed between the verbal and the performance IQs for the first 20 clients. This he attributed to some sort of brain damage. This conclusion was validated by the results of the Rorschach which, although virtually unscorable, contained multiple responses of fire burning in hell, blood oozing from puncture wounds of the face, drowning in icy pools of water, and other dangerous and often implausible statements. Results from the SII were so difficult to interpret that Dr. Rednik did not even try. From these first few test results alone, Dr. Rednik decided to change the focus of his practice from career counseling to evaluating new immigrant clients for social security disability payments. He concluded that all the clients he assessed qualified for disability coverage.

To conclude that these clients were disabled seriously undermines the more obvious conclusion that language and cultural background accounted for the test results. Dr. Rednik did not know the quality of the translation, but even if it were perfect linguistically, the validity of the tests in these translations was certainly problematic. The tests were scored and interpreted inappropriately (Standard 9.02a). They were not validated for use with these populations (Standard 9.02b). Cultural and linguistic differences were not taken into account in the interpretations (Standard 9.06). Above all, Dr. Rednik behaved incompetently. Standard 2.01a, Boundaries of Competence, advises, "Psychologists provide services, teach, and conduct research with populations and in areas only within the boundaries of their competence, based on their education, training, supervised experience, consultation, study, or professional experience." His training and experience in testing and assessment took place years before culture became a central consideration in the assessment process in terms of test selection, test administration, and test interpretation. It could be argued that these deviations from the standard of care were the direct cause

of additional pain and suffering of his clients. Instead of counseling them to find fulfilling employment, Dr. Rednik categorized them for all to see as emotionally disabled and unable to work. A complaint for malpractice would not be inappropriate here.

Test Disclosure and Security

A legal issue both broad and far-reaching for psychological assessment is test security and the release of test records. The Ethics Code, in Standards 9.04 (a) and 9.11, addresses the difference between test data and test materials. Basically, test data include anything that the psychologist or the client adds to the test. These include responses, raw data, and notes. Unlike in years past, these data are considered, in effect, to be in the control of the test taker. The prevailing value is client autonomy, that an individual can make the decisions that affect his or her life, even if the psychologist does not think such an action is in the best interest of the client. Although state law may have something to say, generally if a client or client's legal representative signs a release, the test data must be turned over to the person named in the release. There is some negotiating room in such situations. The psychologist can try to convince the client that test data might be damaging if released, say, to the employer. Campbell, Vasquez, Behnke, and Kinshereff (2010) listed circumstances that the psychologist should make the client aware of with respect to test data release. These include the possible necessity of releasing the complete test results instead of a requested portion, negative consequences such as stigma to disclosure of a particular diagnosis, and risk of records falling into the hands of another party.

When a third party is the client (e.g., in mandated testing by a school), the test taker may not have the legal authority to release the test data. This situation needs to be covered in the informed consent to testing, Ethics Code Standard 9.03. Even if there is a mandate for testing to occur, the test taker still has the ability to refuse to consent to the assessment and submit to the consequences. The limits to confidentiality also need to be explained.

Consider the following example:

> John Archer is a junior psychology major at Pacific Atlantic University. A rash of false fire alarms has been set in several of the university residence halls, necessitating the complete evacuation of the 10-story buildings at 3:00 a.m. The university's hearing board has found that Archer set the alarms. In lieu of expulsion from the university, the dean of students has decided to require a complete psychological assessment of Archer followed by mandated counseling during a 1-year suspension. Archer is told that the assessment results will be sent to the dean and that he will have access to them only with the consent of the dean.

Despite Archer's case being mandated assessment and counseling, he has a choice. He can participate in the assessment and counseling, or he can be expelled from the university followed by possible criminal charges. Nothing about this situation forces Archer to release his test results to another person of his choice without the express permission of the dean, who is the client in this case.

Ethics Code Standard 9.11 instructs the psychologist about what *not* to do with the test materials, and that is to disclose them. Clearly, many psychological tests become less valid and less useful when the test taker knows their content in advance. This is more than an ethical standard: It is also a legal mandate. Most psychological assessment instruments are covered by copyrights. Pursuing copyright violations is important to test publishers who could not stay in business if they were ignored. Even if the copyright to a test is not owned by a test publisher, one should assume that an instrument that is published in a journal at least has a copyright that has been assigned to the journal publisher and that it is important to seek permission before using the test. It should also be noted that photocopying test materials, in lieu of purchasing them, violates federal copyright law.

Courts have had mixed results in honoring the security of tests, thus making it all the more imperative for a psychologist who might be called for a court appearance to know the rules and procedures of the particular court. The issue is that, under Federal Rule of Civil Procedure 26, a party to litigation

must turn over supporting documents to a claim to the opposing party (see Federal Rules of Civil Procedure, 2010). These include the testimony of an expert witness with the information considered in forming the expert opinion. The U.S. Supreme Court has spoken to this issue only once. In *Detroit Edison Co. v. N.L.R.B.* (1979), the Supreme Court found that a court could not require an expert to turn over psychological testing materials without any limitations. In this case, Detroit Edison had tested the aptitude of union employees with the promise of confidentiality. The union demanded access to the testing data and materials on all members during arbitration. The company volunteered to discuss individual outcomes with each union member. The Court held that "the order requiring the Company unconditionally to disclose employee scores to the Union was erroneous" (1979, p. 320).

In *Taylor v. Erna* (2009), the U.S. District Court in Massachusetts dealt with this question. Taylor and her parents sued Erna for brain injury damages resulting from an automobile accident. The parents retained a neuropsychological expert, Thomas Deters, PhD. On the basis of psychological tests, interviews, and affidavits, Dr. Deters wrote a report for the court that did not include test data and materials. Erna objected, demanding the test data and materials. Taylor opposed the request, saying that even though this unconditional disclosure would violate Standard 9.11 of the Ethics Code, Dr. Deters had already given the raw data and materials to Erna's expert, who was also a psychologist. The *Taylor* court noted Ethics Code Standard 1.02, Conflicts Between Ethics and Law, which allows a psychologist to adhere to legal requirements after first making known their commitment to the Ethics Code and then trying to resolve the conflict.[1] In its conclusion, the court noted that courts had gone various ways in limiting test disclosure. The court decided that giving the defendants full access to test data and materials through their qualified experts would be sufficient. In doing so, the court noted that "the apparent conflict between the demands of the Federal Rules and the APA's Ethical Principles has given rise to a number of court-ordered resolutions" (p. 2). In *Tibbs v. Adams* (2009), the court reasoned that a court order to compel production of documents would protect the psychologist from violating the Ethics Code. In *Kayongo-Male v. South Dakota State University* (2008), the court ordered full disclosure to the plaintiff. These cases are solely from U.S. Courts and are meant to show the variability of legal opinions. A particular judge can compel a psychologist to turn over all test materials to a party or protect the psychologist from doing so. There is no substitute for being familiar with the rules of the court to which one's case has been assigned, no substitute for consulting one's own attorney, and no substitute for being highly familiar with the nuances of the Ethics Code.

In addition, familiarity with the requirements of the federal HIPAA is essential. If a client requests that a psychologist release records within their "designated record set," the psychologist is mandated to do so. In the instance where the psychologist feels it would be to the client's disadvantage to have the records released, the psychologist would be wise to consult with an attorney.

Clinical Testing and Assessment in Forensic Contexts

A forensic examination typically takes place when a legal question needs to be answered. A court may appoint a psychologist to make such an assessment or, less often, an attorney contracts with a psychologist for an assessment that is expected to favor the side of the assessed individual. Although this evaluation may often involve the retrospective reconstruction of a client's mental state at the time of an event, such as a will signing or a homicide, this discussion will focus on a more immediate assessment of current functioning. In addition, readers will find information concerning forensic mental health assessment in Chapter 16 of this volume.

Being a Witness

It is important that everyone who practices in the forensic area become familiar with the rules

[1]Since the Taylor case, the wording of Ethics Code Standard 1.02, requiring psychologists to resolve conflicts between the law and the Ethics Code "in keeping with basic principles of human rights." This would not change the discussion of test security.

governing witnesses in their jurisdiction. As Standard 2.01(f) states, "When assuming forensic roles, psychologists are or become reasonably familiar with the judicial or administrative rules governing their roles." If a psychologist is subpoenaed as a witness without prior experience in this role, talking to the client's attorney is one way to become familiar with the procedures of the court but should not be the only way. Attorneys who represent particular clients are decidedly partial. (In fact, attorneys are taught to zealously defend their clients.) One should always be realistically suspicious that an attorney may encourage a treating and/or assessing psychologist to cross the line from a fact witness to an expert witness.

A fact witness is simply someone who states the facts. If a psychologist has evaluated a person, the facts consist of the names of the tests, the results, and the conclusions. The conclusions should be derived directly from the test results modified by the psychologist's notation of personal and situational variables. Clinical judgment should be kept to a minimum. The *Federal Rules of Evidence* (2009) are a good starting point for the definition of expert testimony, although to be sure, child custody evaluations are rarely part of a federal court procedure.[2]

> If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case. (Rule 702: Testimony by Experts)

Whether expert testimony can be admitted at all in a legal case is determined by the trial judge's gatekeeper analysis of the standards specified by the U.S. Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993) and *Kumho Tire Co. v. Carmichael* (1999). These standards superseded the "general acceptance" test that had been set forth in *Frye v. United States* (1923). Summarizing the *Daubert* standards, the Court said,

> The court must therefore make two separate determinations: (1) does the expert's testimony consist of scientific, technical, or other specialized knowledge; and (2) will the application of that knowledge to the particular facts and circumstances of the case aid the jury in understanding or deciding a fact that is in issue. (*Daubert v. Merrell Dow Pharmaceuticals, Inc.* 1993, at 592)

To clarify this point, in an explanation that is particularly relevant to psychological testimony, the Court stated,

> In *Daubert*, this Court held that Federal Rule of Evidence 702 imposes a special obligation on a trial judge to "ensure that any and all scientific testimony . . . is not only relevant but reliable" (509 U.S., at 589). The initial question before us is whether this basic gatekeeping obligation applies only to "scientific" testimony or to all expert testimony. We, like the parties, believe that it applies to all expert testimony. (*Kumho Tire Co. v. Carmichael* 1999, at 147).

What the *Daubert* and *Kumho* holdings mean for the psychologist who becomes an expert witness for either party is that there is a high standard for the reliability and validity of the evidence presented. The testimony must be related closely to the scientific data and then to the issues to be decided by the court.

The Ultimate Issue

Attention also must be paid by the psychologist not to testify to the ultimate issue. The ultimate issue is

[2]Most every child custody evaluation is intended for a case that will be in a state court. The trial-level state court goes by a variety of names but may be the domestic relations court or the juvenile court. Rules of evidence vary from state to state, whereas federal rules are the same. Custody cases in juvenile court are typically those in which the parents are not married or the county has stepped in to protect the children. This also varies by jurisdiction.

the legal question to be decided by the trier of fact (the judge or the jury). The *Federal Rules of Evidence* (2009) speak to this decision:

> (a) Except as provided in subdivision (b), testimony in the form of an opinion or inference otherwise admissible is not objectionable because it embraces an ultimate issue to be decided by the trier of fact. (b) No expert witness testifying with respect to the mental state or condition of a defendant in a criminal case may state an opinion or inference as to whether the defendant did or did not have the mental state or condition constituting an element of the crime charged or of a defense there to. Such ultimate issues are matters for the trier of fact alone. (Rule 704: Opinion on Ultimate Issue)

Melton et al. (2007) concluded that even though ultimate issue testimony is often desired by lawyers and the triers of fact, "mental health professionals ordinarily should refrain from giving opinions as to the ultimate issues" (p. 17). The trier of fact should evaluate the evidence and the testimony of the psychologist and reach their own conclusions. A psychologist who appears in courts as an expert should not simultaneously be a decision maker. This prohibition is considerably mitigated when the psychologist testifies in child custody or child abuse proceedings.

In *United States v. West* (1992) the seventh circuit court, in a case of the insanity defense, analyzed the legislative intent of the Rule 704. Congress gave two reasons for wanting to take the decision from the experts and give it to the jury. The first is described in the Senate Report, which says that "the purpose [of 704(b)] is to eliminate the confusing spectacle of competing expert witnesses testifying to contradictory conclusions as to the ultimate legal issue to be found by the trier of fact." (S. Rep. No. 225, 98th Cong., 1st Sess. 231 [1983], reprinted in 1984 U.S. Code Cong. & Admin. News 3182, 3412). The second rationale Congress gave for 704(b), contained in both the Senate and House Reports, evinces a skepticism not about the spectacle of competing mental health experts and their conflicting testimony but about their competence to testify about moral questions of criminal responsibility. The House Judiciary Committee's report emphasizes that

> while the medical and psychological knowledge of expert witnesses may well provide data that will assist the jury in determining the existence of the defense, no person can be said to have expertise regarding the legal and moral decision involved. Thus, with regard to the ultimate issue, the psychiatrist, psychologist or other similar expert is no more qualified than a lay person. (H.R. Rep. No. 577, 98th Cong. 1st Sess. 2, 16 [1983])

Similarly, the Senate Report, quoting from a statement by the American Psychiatric Association, stressed that

> psychiatrists are experts in medicine, not the law. . . . When "ultimate issue" questions are formulated by the law and put to the expert witness who must then say "yea" or "nay," then the expert witness is required to make a leap in logic. He no longer addresses himself to medical concepts but instead must infer or intuit what is in fact unspeakable, namely, the probable relationship between medical concepts and legal or moral constructs such as free will. (S. Rep., supra, in 1984 U.S. Code Cong. & Admin. News at 3413)

This same conclusion with respect to the ultimate issue has also been applied, for example, in cases of mental retardation and punishment (*Hall v. Quarterman*, 2008), stress, depravity of mind, specific intent (*Haas v. Abrahamson*, 1990), adoption (*Dahlin v. Evangelical Child and Family Agency* (2002), and credibility (*United States v. Sessa*, 1992). Typical instances when ultimate-issue testimony might be allowed involve the behavior and dysfunction of a sexually abused child (*United States v. Palmer*, 1989) and child custody (*Lisa W. v. Seine W.*, 2005). Thus, it would seem that ultimate-issue testimony might be allowed in a case in family court, as an exception to Rule 704. The question of

ultimate-issue testimony is still largely unsettled and varies among courts and jurisdictions (Heilbrun, Grisso, & Goldstein, 2009).

Child Custody Evaluations

Another area of clinical and counseling assessment with important legal implications and potential problems for the practitioner is the evaluation of individual parents for child custody recommendations. Even if a psychologist does not practice directly in this area, it is important to consider the possibility of being subpoenaed as a witness for a client who is in court. This most commonly happens in divorce and child custody cases.

Psychologists who conduct marriage and family therapy must keep in mind that a court appearance is always possible. Thus, it behooves everyone who practices in this area to become familiar with the rules governing witnesses in their jurisdiction, as Standard 2.01(f) states.

A rather typical case is the following (a similar example appears in Gutheil and Hilliard, 2001).

> Marcia Kindsole, PsyD, specializes in family therapy. She has had two sessions with Karen and Timothy Struggles. Before even meeting the children, it has become apparent to Dr. Kindsole that the chemistry is all wrong between her and Mr. Struggles. He glares at her during the sessions and utters snappy and sarcastic comments about her treatment method, making her very uncomfortable and self-conscious with him in the room. She discusses with the Struggles the possibility of continuing therapy with a different psychologist. The result of this conversation is that Karen decides to continue on in individual psychotherapy, whereas Timothy decides to quit therapy all together. In her report to Karen's insurance company, Dr. Kindsole gives Karen a diagnosis of posttraumatic stress disorder (PTSD). From what Karen has said, there is little doubt in Dr. Kindsole's mind that Timothy's demeaning and emotionally hostile treatment of Karen is the cause of her dysfunction. Progress is being made when Karen says that her attorney would like Dr. Kindsole to write a letter for use in her child custody hearing. Upon receipt of the letter, the attorney says he will call Dr. Kindsole as a witness. Dr. Kindsole states to both Karen and her attorney that she can only be a fact witness, relating when she met with Karen for treatment and what she did during that time. Both agree that this is what they want her to testify about. However, things change when the court date arrives. On the witness stand, Dr. Kindsole testifies to her treatment of Ms. Struggles. At this point, Ms. Struggles's attorney asks her for her diagnosis and then for her justification of the PTSD diagnosis. Very skillfully, the attorney leads Dr. Kindsole into a statement that in her opinion, the Struggles's children should be in their mother's custody because of their father's behavior. Now Dr. Kindsole has become both a fact and an expert witness, testifying to her opinion in a child custody case. Dr. Kindsole should have explained to the court that this is not only a multiple relationship but also one that she is not prepared for, and furthermore, that requiring her to testify as an expert would be requiring her to violate her ethics code. She has done no assessment of parenting ability of either party, only of the mother's mental health.

A psychologist who considers becoming involved in child custody evaluations would be well advised to begin with a careful reading of the APA Guidelines for Child Custody Evaluations in Family Law Proceedings (APA Committee on Professional Standards and Practice, 2009), followed by consultation with experts and targeted training. The APA Guidelines represent a revision of the original 1994 Guidelines. The Association of Family and Conciliation Courts's (2006) *Model Standards of Practice for Child Custody Evaluations* should also be consulted for

guidance on the use of psychological tests. Child custody evaluations need to be multifaceted, including a combination of psychological testing, clinical interviews, and observations of behavior. Following Ethics Code Standard 9.01 (b) necessitates an examination of all parties desiring custody along with observation of their interactions with the children. Typically, these individuals have been married or unmarried partners with the best interests of the children being paramount, assuming that the parents are fit. The APA Guidelines note that, because this situation is a contentious, adversarial situation, psychologists need to focus on "skills, deficits, values, and tendencies relevant to parenting attributes and a child's psychological needs" (APA Committee on Professional Standards and Practice, 2009, p. 6). Concentrating narrowly on only those factors directly relevant to parenting helps control the risk inherent in the recommendation of one person over another for custody: "Family law cases involve complex and emotionally charged disputes over highly personal matters, and the parties are often deeply invested in a specific outcome" (APA Committee on Professional Standards and Practice, 2009, p. 8). In Ohio, for example, child custody/domestic relations cases account for 27% of the ethics complaints filed against psychologists with the licensing board (Ronald Ross, personal communication, 2010). Child custody is a treacherous psycholegal practice for the novice, and one in which there is likely a hostile losing party.

A legal case illustrates the danger here.

> Peter Hughes and his wife Pamela Hughes were divorced with two sons. They agreed that the boys would live with Peter. Subsequently, the wife took the sons to live with her, an act that resulted in an emergency conciliation hearing with joint custody awarded. At this time, the judge ordered a full child custody evaluation to be done by employees of the Chester County Custody Evaluation Program. The three mental health professionals involved were a social worker and a master's degree counselor supervised by a psychologist. The report recommended joint custody. Upon learning of the recommendation, Peter sued the three mental health professionals. He alleged that they conspired to deprive him of his primary custody by coaching Pamela prior to the administration of psychological tests, altering test results against him, destroying favorable test results, and misrepresenting significant facts.

Because Peter's claims were not brought under federal law, the judge dismissed the case for lack of jurisdiction (*Hughes v. MacElree,* 1997). It should be noted that, even though this case was dismissed by the federal court, an angry parent would likely refile it in the state court. Defending these cases costs money, time, and often the professional reputation of the psychologist.

A more typical result is from *Burk v. State of Arizona* (2007).

> Angela Burk sued Cathi Culek, a county child custody evaluator, for negligence. It was undisputed that Culek performed her duty in a discriminatory manner by recommending that Angela's former husband be given primary custody of their minor daughter, S.L., including every Sunday because he attended the Church of Jesus Christ of Latter-day Saints. Burk alleged that Culek objected to her "moral choices." Upon Burk's request, the court appointed a second custody evaluator, Dr. Ralph Earle, who disagreed with Ms. Culek's report. After the court granted custody and visitation to Burk according to the schedule she wanted, which followed Earle's recommendation, a still-angry Burk sued Culek for damages.

The judge decided that judicial immunity extended to Culek because she had been appointed by the court. Many courts have held that in evaluating child custody as a result of court appointment, psychologists are immune from damages because

their evaluation and report is a part of the judicial process. Often, a psychologist involved in a child custody assessment can ask the judge to be court appointed.

According to Otto and Edens (2003), most states have adopted the child custody standards listed in the Uniform Marriage and Divorce Act (1987). The standards consider the wishes of the parents; the wishes of the children; the relationships between the children and their siblings and other people who are good for them to be with; the child's adjustment to home, school, and community; and the physical and mental health of all concerned. These standards are left to be further defined by legislatures and judges. The first four of these factors can be estimated by behavioral observations and clinical interviews. The final factor, mental health, should be assessed through a combination of observations, interviews, and psychological testing. According to Melton, Petrila, Poythress, and Slobogin (2007), the number of cases where a mental health professional is involved is less than 10% of the 10% of custody cases decided by a judge. Most cases are decided by the parents, sometimes with the help of mediation or attorney negotiations. Sometimes one party will hire a psychologist to evaluate the parents for custody, but when this happens, the other party may not cooperate. At other times, if the parties are particularly intransigent, the court will appoint an independent family evaluator with whom both parties are ordered to cooperate.

Some would argue that in light of the five aforementioned factors, psychological testing is of limited value in child custody evaluations; however, most evaluators do use assessment instruments when determining the mental health of the parents (Melton et al., 2007). The most commonly used test is the MMPI–2, a self-report measure described more fully in Chapter 11 of this volume, which also touches on forensic applications of the test. Several factors should be considered when using this test. The first is that the evaluation of mental health in such a manner must be directly relevant to the custody decision to be made. The second is that the results of the test must be interpreted in light of the current mental state of the person being assessed. In their casebook on responsible test use, Eyde, Robertson, and Krug (2010) used an example of giving the MMPI–2 to a mother for a child custody evaluation. Among other results, the woman had an elevated Scale 6 score, indicating that reports of some of the problems she had with her husband might well be delusions. On further evaluation, it was determined that this was a realistic response to threatening and bizarre behavior on his part. As indicated in Ethics Code Standard 9.06, it is important when interpreting test results to consider situational and other variables that might reduce the accuracy of the interpretation.

Many attempts have been made to develop specialized tests for use in child custody evaluations. Surely it would be helpful to have an instrument that would measure parenting capacity more precisely, rather than require the psychologist to infer parenting ability or the quality of the parent–child relationship from measures of personality and mental health. Authors such as Melton et al. (2007) and Otto and Edens (2003) have concluded that reliability and validity evidence for these more specific instruments is often inadequate. Otto, Edens, and Barcus (2000) described child custody evaluations as "the most complex and difficult type of forensic evaluation" (p. 312). These instruments include but are not limited to the Ackerman–Schoendorf Scales for Parent Evaluation of Custody (ASPECT; Ackerman & Schoendorf, 1992) and the Bricklin Perceptual Scales (BPS; Bricklin, 1990) for children.

Assessing Competencies and Damages

Psychologists are frequently called upon to assess pain and suffering and emotional injury in tort cases. In addition, there is a wide array of competencies in forensic cases, all of which require a formal psychological assessment. These include, but are not limited to, competence to stand trial, testamentary capacity, insanity, and competence to make medical decisions.

Tort Damages

A *tort* is a type of legal case that involves a civil, as opposed to criminal, injury. These cases include many types of claims such as negligence, defamation,

trespass, and malpractice. One type of injury is emotional/mental pain and suffering. An illustrative example follows.

> Amy Starkweather was under the treatment of Janet Fellows, PhD, for the adult effects of an abusive childhood, attachment issues, and intermittent depression. Dr. Fellows persuaded Amy that having a sexual relationship with her would be helpful in her recovery. Thirteen years later, a shattered Amy emerged from this relationship with the knowledge that she was feeling worse than ever. Dr. Fellows had become the center of her life. She felt used and betrayed. She sued Dr. Fellows for malpractice, alleging that she abandoned her duty of care and exploited her client. Amy alleged that she sustained financial loses as well as grave mental injuries, including an aggravation of her preexisting condition, as a result of her negligent treatment by Dr. Fellows. She has contracted with Holly Thorndike, PsyD, to be her mental health evaluator/expert.
>
> In addition to interviews, inspection of cards and letters from Dr. Fellows to Amy, and examination of medical records, Dr. Thorndike was delighted to find a battery of psychological tests given to Amy during her first few months of therapy. (This is a critical element because if an individual has a preexisting condition, it is difficult to prove one of the elements of negligence—that the negligent act was the proximate cause of the injury.) The instruments that had been used were the MMPI–2, the WAIS–III, and the Beck Depression Inventory—II. Dr. Thorndike wrote that it was clear that Amy was and is still acutely depressed and that Dr. Fellows's psychotherapy was nonviable. Since the ending of the relationship she continually was suffering from panic attacks, thoughts of death, and anxiety. As a result of this relationship her preexisting condition has steadily deteriorated and she has sustained severe emotional damage. She has been hospitalized four times. Despite her intelligence, she has been unable to work. She is distrustful of forming social relationships. Her depression has deepened, and she has twice attempted suicide. She is enraged over her victimization and terrified at the thought that this may have been her fault.

The question now is how to show that Amy's condition is the direct result of her negligent treatment. The opposing attorney would make a big point of the fact that Amy was not in good mental health to begin with and that other life events and circumstances may well have caused this deterioration. The behavior of Dr. Fellows is unrelated to Amy's current functioning, the attorney argues. The finder of fact, be it judge or jury, would make the final causal determination. The job of the psychologist is to show pre- and postmental functioning where the only logical causal conclusion is the defendant's behavior. This is the ultimate issue.

Melton et al. (2007) noted three points in writing such a report for the legal system. The first is that the report should go well beyond a diagnosis to citing the evidence for the mental injury. Second, "a complete assessment of mental injury requires a longitudinal history of the impairment, its treatment, and attempts at rehabilitation, including the claimant's motivation to recover" (p. 422). Third, the conclusion should be left to the legal decision maker.

Assessing Civil Competencies

A forensic psychologist typically assesses competencies in the criminal and civil law. Criminal competencies include competency to stand trial, to plead guilty, and to waive Miranda rights. These are usually testified to in a court of law. Civil competencies include testamentary competence and the competence to make medical decisions. Forensic psychology traditionally has been a postdoctoral specialty, but, more frequently, psychologists have been tempted to assess competencies with the motivation

that to do so provides an income that is not subject to the whims of managed care organizations. Training through continuing education courses and consultation with a colleague may suffice. Before undertaking the first assessment of competency, a psychologist should secure the necessary training and read thoroughly and thoughtfully the Division 41 Specialty Guidelines for Forensic Psychologists (Committee on Ethical Guidelines for Forensic Psychologists, 1991). This document, which is considered to be aspirational and nonenforceable, is most helpful to those who interact with the legal system only occasionally. Consistent with the Ethics Code, it emphasizes competence, limits of confidentiality, informed consent (unless court ordered), and attention to the possibility of a multiple relationship.

Competence to make treatment decisions. Generally, the issue of competence to make treatment decisions arises when an individual refuses treatment that can potentially save a life or restore quality of life. How mentally handicapped can a person be and still be competent to make a medical decision? This is also relevant when the decision maker is a minor. Courts frequently have enabled adolescents over the age of 15, or "mature minors," to make medical decisions regarding sexual or reproductive health or substance abuse treatment.

For example,

> Martin Jepson is a 16-year-old boy newly diagnosed with neurofibromatosis, a disfiguring condition of tumors on his face. As a result, he has lived as a recluse with only his mother to care for him and to socialize with him. He dropped out of school at the age of 14 when his facial deformity was apparently severe enough to prompt teasing and bullying from his classmates. Surgery would improve his facial appearance decidedly and permit him to leave his home, finish school, and enter the workforce. The surgery would result in a painful recovery and probably will need to be repeated, depending on the regrowth of his tumors. Martin considers the two options and decides to have the surgery. His mother, however, does not want him to proceed with this plan. She is afraid he might die in surgery or somehow come out of it in worse shape than he is currently. The consulting psychologist believes that Martin's mother has found her purpose in life to care for her deformed child and will feel displaced if he is to live a more normal life. The presumption that Martin is incompetent to make this decision because he is a minor can be rebutted by a psychologist who uses a clinical interview along with a structured diagnostic assessment.

Grisso (2003) identified four abilities that contribute to form a rebuttal of the presumption of incompetence: to understand the relevant information, to appreciate the relevance of the information, to use the information to make the decision, and to communicate the decision (p. 395). The MacArthur Competence Assessment Tool for Treatment (MacCAT-T; Grisso & Appelbaum, 1998) includes questions about situations that deal with understanding, appreciation, reasoning, and choice. A hypothetical example of a structured interview question might be as follows:

> Think of your symptoms. Often, people with your type of emotional problem can be helped. Your doctor might prescribe medication or ask you to talk to a trained mental health professional. Do you feel that either of these treatments might benefit you?

Similar questions could be constructed regarding treatment for a physical disorder.

Testamentary competence. Another area of civil competency that is an interesting one for a psychologist to assess is the competence to execute a will, or *testamentary* competence. For a will to be valid, a testator must be of sound mind and under no undue influence. The standard for testamentary competence was first stated in the English case of *Banks v. Goodfellow* (1870). It is surprising that this test is still used today. The testator must know (a) that he or she is making a will, (b) the extent of his or her property, (c) the "natural objects of his or her bounty," and

(d) how the will shall be used to distribute the property. What is meant by the "natural objects of one's bounty?" These are presumed to be family members or long-time close friends of the testator.

In many instances, testamentary competence is assessed retroactively, that is, after the testator is deceased. However, this situation is not always the case. Melton et al. (2007, p. 398) advised taking each element of testamentary competence separately during assessment. In addition to a general clinical evaluation of the testator's mental health, a psychologist should question the individual on the meaning and purpose of a will and, particularly, why the individual is making or changing a will at this time. Critical areas of probing include what testators know about their property and how they have made the decision to bequest particular property to particular individuals. This is especially important if the state standards for intestate succession are not approximated. For example, if one child is left out of an inheritance, why was this decision made? If a caretaker or an assisted living facility is chosen to inherit a large portion of the estate, why was this decision made? Was there any possibility of undue influence or coercion? Melton et al. suggested that general questions about property include a person's former or present employment, salary, and type of investments. Questions about the natural objects of one's bounty would include a discussion of family members and other individuals who have been significant figures in a person's life. A discussion of each specific bequest also may be important.

Although assessment of testamentary capacity generally takes the form of a structured clinical interview covering the legal elements, the use of one or more psychological tests is not out of order. If a testator is judged to be incompetent and/or under undue influence, a judge will invalidate the will. In this instance, either a prior will may be substituted or the laws of intestate succession determine the inheritance of the property.

CONCLUSION

Many of the legal issues in clinical and counseling testing and assessment can be traced back to the basic ethics of the profession. The first part of this chapter focused on malpractice as it results from deviations from the professional standard of care. The standard of care for psychological assessment in many states is found in the Ethics Code (APA, 2010). In other states, the duty of care of psychologists is codified and reads very closely to the APA Ethics Code. Violations of the standards having to do with confidentiality, competence, and informed consent are frequent breaches of the duty of care that can cause some sort of psychological injury to clients or patients and for which they can sue for malpractice damages.

Any assessment activity likely to result in a court appearance, whether as a fact or an expert witness, has its own risk management strategies; the most important one is being familiar with the rules and procedures of the court and avoiding the perils of multiple relationships. The second part of the chapter discussed legal issues arising from assessments by psychologists that are directed toward making a legal decision. The most common of these is the assessment of a competency (e.g., to stand trial, to plead guilty, to act as one's own attorney, to write a will), release of test data, and assessment of parenting capacity that becomes part of a child custody decision.

References

Ackerman, M., & Schoendorf, K. (1992). *ASPECT: Ackerman–Schoendorf Scales for Parent Evaluation of Custody—Manual.* Los Angeles, CA: Western Psychological Services.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010).* Retrieved from http://www.apa.org/ethics/code/index.aspx

American Psychological Association Committee on Professional Practice Standards. (1994). Guidelines for child custody evaluations in divorce proceedings. *American Psychologist, 49*, 677–680. doi:10.1037/0003-066X.49.7.677

American Psychological Association Committee on Professional Practice and Standards. (2009). *Guidelines for child custody evaluations in family law proceedings.* Retrieved from http://www.apapracticecentral.org/news/2009/child-custody.aspx

Association of Family and Conciliation Courts. (2006). *Model standards of practice for child custody evaluations.* Retrieved from http://www.afccnet.org/resources/standards_practice.asp

Banks v. Goodfellow, 5 Q. B. 549 (1870).

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). New York, NY: Oxford University Press.

Bricklin, B. (1990). *Bricklin Perceptual Scales manual.* Furlong, PA: Village.

Burk v. State of Arizona, 156 P. 3d 423 (Ct. of Appeals of Ariz. 2007).

Campbell, L., Vasquez, M., Behnke, S., & Kinscherff, R. (2010). *APA Ethics Code commentary and case illustrations.* Washington, DC: American Psychological Association.

Civil Rights Act, Public Law 82–352 (78 Stat. 241) (1964).

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty guidelines for forensic psychologists. *Law and Human Behavior, 15*, 655–665. doi:10.1007/BF01065858

Dahlin v. Evangelical Child and Family Agency, NO. 01 C 1182, 2002 U.S. Dist. LEXIS 24558 (N. D. of Ill. Dec. 18, 2002).

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 573 (1993).

Detroit Edison Co. v. N. L. R. B., 440 U.S. 301 (1979).

Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior* (2nd ed.). Washington, DC: American Psychological Association.

Family Educational Rights and Privacy Act (FERPA), 20 U.S. C. § 1232g; 34 C. F. R. pt. 99 (1974).

Federal Rules of Civil Procedure. (2010). Eagan, MN: West.

Federal Rules of Evidence. (2009). Eagan, MN: West.

Fisher, C. B. (2003). *Decoding the ethics code: A practical guide for psychologists.* Thousand Oaks, CA: Sage.

Frye v. United States, 293 F. 1013 (DC Cir 1923).

Griggs v. Duke Power, 401 U.S. 424 (1971).

Grisso, T. (Ed.). (2003). *Evaluating competencies: Forensic assessments and instruments* (2nd ed.). New York, NY: Kluwer Press.

Grisso, T., & Appelbaum, P. S. (1998). *MacArthur Competence Assessment Tool for Treatment (MacCAT-T).* Sarasota, FL: Professional Resource Press.

Grisso, T., Appelbaum, P. S., Mulvey, E. P., & Fletcher, K. (1995). The MacArthur Treatment Competence Study II: Measures of abilities related to competence to consent to treatment. *Law and Human Behavior, 19*, 127–148.

Gutheil, T. G. (2009). *The psychiatrist as expert witness* (2nd ed.). Washington, DC: American Psychiatric Press.

Gutheil, T. G., & Hilliard, J. T. (2001). The treating psychiatrist thrust into the role of expert witness. *Psychiatric Services, 52*, 1526–1527. doi:10.1176/appi.ps.52.11.1526

Haas v. Abrahamson, 910 F. 2d 384 (7th Cir. 1990).

Hall v. Quarterman. 534 F. 3d 365 (7th Cir. 2008).

Health Insurance Portability and Accountability Act (HIPAA) 29 U.S. C. 1181 (1996).

Heilbrun, K., Grisso, T., & Goldstein, A. M. (2009). *Foundations of forensic mental health assessment.* New York, NY: Oxford University Press.

Hughes v. MacElwee, No. CIV-A. 97–3304, 1997 U.S. Dist. LEXIS 18069, (E. D. Pa. Nov. 13, 1997).

Kayongo-Male v. S. D. State Univ., No. CIV 04–4172, 2008 U.S. Dist. Lexis 51602 (Dist S. D. July 3, 2008).

Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

Larry P. v. Riles, 793 F. 2d 969 (9th Cir. 1984).

Lisa, W. v. Seine W., 862 N. Y. S. 2d 809, Family Ct. Kings Cty. NY. (2005).

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (2007). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (3rd ed.). New York, NY: Guilford Press.

Otto, R. K., & Edens, J. F. (2003). Parenting capacity. In T. Grisso (Ed.), *Evaluating competencies: Forensic assessments and instruments* (2nd ed., pp. 229–308). New York, NY: Kluwer.

Otto, R. K., Edens, J. F., & Barcus, E. H. (2000). The use of psychological testing in child custody evaluations. *Family Court Review, 38*, 312–340. doi:10.1111/j.174-1617.2000.tb00578.x

Taylor v. Erna, No. CIV 08–10534-DPW, 2009 U.S. Dist. Lexis 69033 (Dist. Mass., August 3, 2009).

Tibbs v. Adams, No. CIV S-05–2334 LKK KJM P, 2009 U.S. Dist. Lexis 90551 (E. D. Cal., Sept. 29, 2009).

Uniform Marriage and Divorce Act, §*402. Uniform Laws Annotated*, *9A*, 561 (1987).

United States v. Palmer, 29 M. J. 929 (Air Force Court of Military Review, 1989).

United States v. Sessa, 806 F. Supp. 1063 (E. D. N. Y. 1992).

United States v. West, 962 F. 2d 1243 (7th Cir. 1992).

PART II

CLINICAL AND HEALTH PSYCHOLOGY

CHAPTER 7

THE CLINICAL INTERVIEW

Katie L. Sharp, Alexander J. Williams, Kathleen T. Rhyner, and Stephen S. Ilardi

A half century ago, influential American psychiatrist Harry Stack Sullivan characterized the clinical interview as

> a situation of primarily vocal communication . . . for the purpose of elucidating characteristic patterns of living of the subject person, the patient or client, which patterns he experiences as particularly troublesome or especially valuable, and in the revealing of which he expects to derive benefit. (Sullivan, 1954, p. 4)

In effect, the interview is a form of conversation with an explicit clinical purpose.

The practice of clinical interviewing has long been regarded as a foundational element within the disciplines of both psychiatry and clinical psychology (Groth-Marnat, 2009; Sullivan, 1954; Wiens, 1976). It was the first systematic method of collecting pertinent patient information in mental health settings. In fact, the ability to conduct a successful clinical interview remains an essential skill for contemporary mental health professionals, one integral to the related processes of assessment and intervention (Nuttall & Ivey, 1986; more extensive information on the assessment process appears in Chapter 2 of this volume).

Clinical interviews have evolved over the past century into a variety of forms, which may be broadly subdivided into three main categories:

- *structured interviews*, in which the wording and ordering of each query is explicitly specified (*semistructured interviews* constitute a somewhat more flexible variation on the theme);
- *unstructured interviews*, which follow a much more open, free-flowing form, one only minimally directed by the clinician; and
- *therapeutic interviews*, in which the interview process is explicitly regarded as a mode of intervention in its own right.

However, before turning attention to an examination of each major interview modality, it may be instructive first to look briefly at the relevant history of the clinical interview.

HISTORICAL CONSIDERATIONS

The information-gathering interview with mentally ill individuals and their families is a practice of ancient provenance. It is described, for example, in the work of pioneering 18th-century clinicians such as Benjamin Rush and Philippe Pinel (Whitaker, 2010). Such conversations were often modeled after the directive question-and-answer format characteristic of other medical disciplines. In the early 20th century, however, as psychiatry and clinical psychology coalesced into distinctive disciplines, Sigmund Freud championed an alternative, psychoanalytically inspired approach to eliciting relevant clinical material: a free-flowing, supple, minimally structured interview format (Groth-Marnat, 2009). Although Freud still viewed the interviewer as an active, dynamic part of the information-gathering process (Nuttall & Ivey, 1986) and a vital player in the clinical dialogue (Matarazzo, 1965), he regarded the unstructured interview as the context in which the patient's important unconscious mental processes might best become evident.

DOI: 10.1037/14048-007
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

Freud's seminal influence on the field endured for decades, and by midcentury his generally nondirective, unstructured approach was the dominant interview modality in mental health settings—one that even found some support from nascent research in the area. An early study of recorded interview protocols, for example, indicated that nondirective interactions frequently preceded favorable changes in client attitudes, in contrast with directive approaches, which often led to client hostility and defensiveness (Snyder, 1945). Likewise, Porter (1950) noted the generally salubrious effects of more open and reassuring, as opposed to evaluative, interviewer responses.

Although trained in a broadly psychoanalytic framework, Sullivan (1954) introduced an influential shift in perspective on the interview process. He conceptualized the psychiatric interview as a *system* of interpersonal processes, one in which there is continuous reciprocal influence between the interviewer and interviewee, with each participant continuously responding to (and influencing) the emotional tone of the other. Sullivan referred to the process as "reciprocal emotion," which held the potential, he averred, to benefit the interviewee through the interviewer's influence. In this respect, Sullivan presaged subsequent therapeutic developments in which the interview process came to be regarded as a potent form of intervention in its own right—as, for example, in the formulation of *motivational interviewing* (Miller & Rollnick, 2002) or even Rogers's (1961) *client-centered* therapeutic approach.

Sullivan's emphasis on the reciprocal causal dynamic between clinician and client was also reflected in the subsequent development of *general systems theory* (Matarazzo, 1965; Watzlawick, Beavin, & Jackson, 1966), which focused on the interactions within the interviewer–interviewee dyad, rather than the actions of each individual alone. Numerous family therapy interventions likewise originated from these dynamic interactional emphases (Craig, 2003; Van Dyke, 2005; for information about assessment in marriage and family counseling contexts, see Chapter 33 of this volume).

In a rather striking contrast, however, behaviorally oriented clinicians (including those subsumed by the *cognitive–behavioral* rubric) have, for the past several decades, focused on the advantages of tightly *structuring* the interview process so that it will reliably elicit valid clinical information of high salience to the processes of both case conceptualization (diagnosis) and treatment (Craig, 2003). Lazarus (1973), for example, proffered a mnemonic ("BASIC ID") to delineate a broad array of topics to be systematically queried by the clinician during an initial interview:

B = behavior, especially problem behaviors
A = affective responses, especially harmful ones
S = sensory alterations (e.g., emotional pain)
I = imagery (e.g., fantasies)
C = cognition, especially distorted thinking
I = interpersonal relationships
D = drugs (both illicit and prescribed)

Likewise, Wolpe (1973) advised behavioral clinicians to collect structured interview-based data concerning early family history, education, employment, sexual history, and current relationships, whereas Kanfer and Scheft (1988) advocated a thorough investigation of the presenting problem; the cause(s) of abnormal behavior; relevant motivation and development; self-control skills; social relationships; and the client's social, cultural, and physical environment.

As the cognitive–behavioral perspective has emerged as an increasingly influential conceptual framework within clinical psychology and related disciplines from the 1960s to the present, the field has witnessed a corresponding increase in the use of structured interviewing. However, another important impetus comes from the growing recognition within psychiatry and allied fields that structured interviews are well suited to providing reliable and valid clinical information for the purposes of psychiatric diagnosis based on criteria of the *Diagnostic and Statistical Manual of Mental Disorders* ([4th ed., text revision; *DSM–IV–TR*; American Psychiatric Association, 2000]; Shea, 1990)—information of great importance both to practicing clinicians (who are routinely required to provide accurate diagnostic information to insurers) and to clinical researchers.

In summary, the historical development of the clinical interview is a progression inextricably interwoven with its diverse conceptualizations. It is not surprising that clinical interviews have assumed a

variety of forms, which generally reflect, more than anything else, the theoretical vantage points, clinical goals, and values of those conducting the interview.

THE UNSTRUCTURED CLINICAL INTERVIEW

An *unstructured clinical interview* involves an open-ended, free-flowing interaction between the clinician and client/patient. It is characterized by the absence of a predetermined set of questions, with a focus instead on spontaneous content that emerges during the clinical interaction. The role of the clinician is one of facilitating the discussion of topics initiated by the interviewee rather than directing them through a structured series of questions solely of the interviewer's choosing (Ilardi & Branstetter, 2004). Indeed, the interviewer generally proceeds under the assumption that important issues will spontaneously arise within the conversation as long as the client is not hindered by the clinician's preconceptions, perspectives, or assumptions (Jenkins, 2007); yet the clinician remains active and engaged, inasmuch as the process relies on "the empathically attuned and dynamically sophisticated clinician given free rein to practice his or her craft" (Shedler, 2002, p. 433).

The clinician conducting an unstructured interview is also free to gently steer the conversation to facilitate assessment of whatever constructs she or he feels is necessary, using whatever responses, questions, or observations she or he believes to be most relevant (Widiger, 2008). In fact, most interviewers proceed with a general set of content domains about which they are interested in gathering information—areas such as the presenting problem, diagnostic status, available social support, and relevant psychosocial stressors (Ilardi & Branstetter, 2004). The loose structure of the unstructured interview also provides considerable freedom for the clinician to pursue any relevant topic or line of inquiry that emerges spontaneously during the course of the interview (Johnson, 1981).

It is important to note that the unstructured interview format may also impel the clinician to attend carefully to each issue brought up by the client, a process with clear potential to strengthen the therapeutic alliance. This consideration, perhaps more than any other, may explain why the unstructured interview remains the most commonly used form of clinical assessment in contemporary practice, despite the well-established psychometric advantages of structured and semistructured approaches (Craig, 2003; Miller, 2003; Sommers-Flanagan & Sommers-Flanagan, 2003).

The unstructured interview, as noted previously, has its roots in psychoanalytic practice. Within the classic psychoanalytic framework, such a nondirective interview is believed to serve as a catalyst for the patient's expression of unconscious material, for example, by means of transference reactions and free associations. Within contemporary practice, however, the unstructured interview is used by clinicians from a wide range of theoretical backgrounds—not just psychodynamic—although the clinician's theoretical orientation may still influence the focus of the interview and the specific details that are emphasized and pursued (Ilardi & Branstetter, 2004).

One implication of having no predetermined structure of questions is that unstructured interviewing relies heavily on the clinician's therapeutic acumen and creativity (Mohr & Beutler, 2003). As a result, unstructured interviewing is a rather difficult skill to acquire, one that requires considerable training and practice to develop (Carlat, 2005). Therefore, most training programs devote a great deal of attention to helping students develop the therapeutic listening skills necessary to be effective in such an unstructured interview context. When skillfully implemented, the interview takes on a natural flow and typically remains focused on the issues presented by the client and the relationship between the client and the interviewer (Johnson, 1981).

The unstructured interview process has been hypothesized by Sullivan to consist of four primary stages, delineated in his influential book, *The Clinical Interview* (1954). The *inception* includes welcoming the patient and establishing the expectations for the interview. *Reconnaissance* consists of questioning the patient about his or her history, social situation, and therapeutic needs. During the *inquiry* stage, the clinician begins to test different clinical hypotheses related to the patient's presenting problems.

The final stage, *termination*, consists of ending the interview in a mutually satisfactory fashion.

A more recent description of the unstructured interview is found in Shea's (1988) five-phase structural model. Phase One consists of the introduction between the patient and therapist, during which the patient is educated about the interview process. Phase Two involves the patient's own account of the presenting problem(s). Phase Three denotes the body of the interview, wherein the clinician attempts to gather more information about the presenting problem as well as related content pertinent to diagnosis and treatment planning. Phase Four consists of summarizing the interview and presenting the clinician's current conceptualization of the patient's problems and how they may be addressed. Termination, the final phase, involves the formal conclusion of the interview and the exiting of the patient.

> Shea's model nicely illustrates the point that although the interview format may remain unstructured (inasmuch as the questions and wording thereof are not planned prior to the session), the interviewer typically retains a general idea of the sequence and flow of the interview. (Ilardi & Branstetter, 2004, p. 1018)

Unfortunately, unstructured interviewing appears be among the least reliable and valid of extant assessment procedures (Maruish, 2008). There is also scant published research concerning its overall clinical utility, but what little there is suggests that it may frequently be inferior to structured and semistructured interviews on that front as well (Basco, 2003; Segal, Hersen, & Van Hasselt, 1994; Vitiello, Malone, Buschle, & Delaney, 1990; Widiger, Sanderson, & Warner, 1986; Young, O'Brien, Gutterman, & Cohen, 1987). Unstructured interviews even show poor psychometric properties regarding the assessment of a patient's relevant psychiatric social history (Ferriter, 1993) and treatment expectations (Ruggeri, Dall'Agnola, Agostini, & Bisoffi, 1994).

Notably, unstructured interviews may also inadvertently introduce substantial interviewer biases regarding patient assessment—a phenomenon to which structured interviews appear to be considerably less vulnerable (Groth-Marnat, 2003). For example, a client who is courteous and polite during the interview may be assessed, via the *halo effect*, as more competent and interpersonally effective than they actually are (Groth-Marnat, 2009). Likewise, the interviewer may be guided by his or her early inferences about a client to probe preferentially for information that confirms such preconceptions, a process known as *confirmation bias*. Such biases may, of course, compromise the reliability and validity of interview-based inferences.

Diagnostic variability in the unstructured assessment context is also of some concern. It often stems from *information variance*, in which the clinician obtains different information from each patient because of wide variation in the range of symptom-related material actually covered during each interview. Diagnostic differences can also be affected by *criterion variance*, in which discrepancies emerge regarding the criteria used to make a decision on the presence or absence of a specific condition. Finally, interviewers vary with respect to their level of experience, theoretical orientation, and disposition, all of which may contribute to the observed variance in assessment outcomes based on the unstructured interview (Rosqvist, Bjorgvinsson, & Davidson, 2007). It is not surprising that the aforementioned sources of variability inherent in the unstructured format generally result in lowered diagnostic reliability and validity in comparison with formal structured interviews (Huffcutt & Arthur, 1994; Marchese & Muchinsky, 1993).

On the other hand, many clinicians still champion the unstructured format. It has been suggested, for example, that information derived from the unstructured interview is particularly useful as a means of augmenting data from other modes of assessment (Bagby, Wild, & Turner, 2003). Likewise, although clinical diagnostic interviews inevitably involve the goal of making accurate *DSM–IV–TR* (American Psychiatric Association, 2000) diagnoses (Jones, 2010), a less rigidly structured interview, in the hands of a skilled diagnostician, may provide the clinician with optimal flexibility in reaching this central assessment goal (O'Brien & Tabaczynski, 2007). Additionally, the client who is presented

with broad, open-ended questions may be especially likely to report his or her spontaneous thoughts, memories, and feelings, thereby to reveal key elements of his interpersonal style (Jenkins, 2007)—potentially pertinent information that could be missed with more structured approaches that utilize a fixed form of questioning.

The unstructured approach may also provide the clinician with a better framework than the structured interview context for establishing an effective rapport with the client, inasmuch as the therapist–client interaction is not limited and constrained by an inflexible set of standardized questions (Jones, 2010). Likewise, the more natural, conversational feel of the unstructured approach could plausibly give the interviewee greater latitude to be open about his or her concerns and to feel more comfortable with spontaneous interjections that could be of clinical significance.

In summary, the unstructured interview remains a widely utilized clinical practice, largely on the basis of its perceived, although empirically unsubstantiated, advantages. Despite its widespread usage, the psychometric soundness of unstructured interviewing remains in question, and its actual clinical utility is also unclear. However, in light of the rather scant amount of direct evidence available, and the many hypothesized advantages of the format, further investigation is necessary before any sweeping negative conclusions regarding unstructured interviewing are warranted.

THE STRUCTURED INTERVIEW

The structured interview is characterized by a predetermined set of queries that the clinician is directed to ask the patient verbatim, in a precisely defined order (Hong & Ilardi, 2004). It also utilizes tightly operationalized and standardized criteria for the coding, scoring, and interpretation of each interviewee response (Beutler, 1995). Because no deviation is permitted from one clinician to the next on any major aspect of the interview process—the phrasing of questions, the temporal ordering of queries, or the discussion of any new or unexpected material presented by the interviewee—the process is designed to elicit nearly identical interviewee responses across different interviewers, thereby to promote heightened interrater reliability (Segal, 1997).

The semistructured interview is an important variation on the structured interview theme. It retains the major elements of the structured approach but differs by virtue of allowing the clinician some latitude in formulating his or her own follow-up queries to further probe relevant content domains (e.g., specific diagnostic criteria, patient mental status, psychosocial functioning, symptom severity) in sufficient depth to permit a valid rating (Hong & Ilardi, 2004). For example, a clinician might attempt to clarify the clinical significance of a patient's self-reported social withdrawal with a follow-up query such as, "You said a moment ago that you haven't been as interested in spending time with others lately: Can you give me an example of one of your usual social activities that you decided not to take part in this week?"

It is important to note, however, that semistructured interviews typically outline for the interviewer most of the potential follow-up prompts to be used in probing for additional clarifying information within each content domain of interest; for example, "When did you first begin to experience difficulty with _____?" The actual degree of interviewer latitude is therefore quite narrow in practice. Accordingly, the semistructured interview may best be regarded as merely a more nuanced version of the structured interview format rather than as a distinct category of interviewing in its own right.

THE SHIFT TO STRUCTURED INTERVIEWING

Structured interviews have become much more widely utilized in both applied and research settings in recent decades. In part, the shift reflects the field's increasing emphasis on *DSM*-based diagnosis, in tandem with a burgeoning recognition that diagnoses derived from nonstandardized, unstructured, idiosyncratic interview procedures exhibit generally poor reliability, with unacceptably high diagnostic variability from one clinician to the next (Helzer et al., 1977).

In contrast, structured interviews allow the clinician (or clinical researcher) to systematically assess every symptom domain within all relevant diagnostic categories (Shea, 1990) using standardized queries and well-operationalized coding criteria – a process that reduces both criterion variance (the use of different standards by different interviewers in determining the presence or absence of each symptom) and information variance (variation in the actual clinical data obtained by each interviewer). Accordingly, structured diagnostic interviews typically yield a high degree of interrater reliability (Segal, 1997; Segal & Falk, 1998). Also, inasmuch as an instrument's reliability serves as a rate-limiting factor on its potential validity—that is, a measure cannot be valid if it is not reliable—structured interviewing is, not surprisingly, associated with enhanced diagnostic validity in comparison with unstructured approaches (Hersen & Bellack, 1988).

In light of such superior psychometric properties, structured diagnostic interviews have become the de facto gold standard for *DSM*-based diagnostic assessment in research contexts. In fact, their inclusion in study protocols may be regarded as virtually a prerequisite for the publication of clinical research on any well-defined Axis I or Axis II population (Hong & Ilardi, 2004).

Nevertheless, as intimated in the preceding section, structured interviews are also characterized by a few attendant limitations. Notably, because the precise phrasing and ordering of all queries are rigidly predetermined, the structured interview format provides little freedom for the skilled interviewer to tailor the process to optimally suit the needs of any particular interviewee or even to give the interview a more natural conversational feel. As a result, both clinician and patient may, at times, find the structured interview format to be somewhat interpersonally awkward (Beutler, 1995). Additionally, because some interviewees may be less willing to divulge sensitive personal information to clinicians with whom they have not yet established a comfortable rapport, the rigidity of the structured interview format may inadvertently induce some patients to withhold critical information that they might otherwise be willing to disclose in a less structured context (Hong & Ilardi, 2004).

COMMONLY USED STRUCTURED INTERVIEWS

Structured interviews are widely utilized in contemporary practice, not only for the purposes of *DSM*-based diagnosis but also for a wide range of clinical assessment purposes, including the evaluation of mental status, symptom severity, global functioning, and relevant social support. The following is a review of five commonly used interviews.

Structured Clinical Interview for *DSM–IV* (SCID)

The SCID is a semistructured interviewing measure utilized in more than 1,000 published clinical research studies (Summerfeldt, Kloosterman, & Antony, 2010). Because clinicians and researchers often have somewhat different diagnostic assessment needs, there are two major editions of the SCID: the clinical version of the SCID (SCID-CV; First, Spitzer, Gibbon, & Williams, 1996) and the research version of the SCID (SCID-I; First, Spitzer, Williams, & Gibbon, 1995). The SCID-CV is the shorter form. It assesses only the disorders most commonly encountered in practice settings (e.g., major depressive disorder, bipolar disorder, posttraumatic stress disorder [PTSD]), and provides abbreviated versions of the Substance Use Disorder and Mood modules. The SCID-I comes in three forms. The patient edition of the SCID-I (SCID-I/P; First, Spitzer, Gibbon, & Williams, 1997c) is the broadest version, designed for research participants who have already been designated as psychiatric patients. The SCID-I/P with Psychotic Screen (First, Spitzer, Gibbon, & Williams, 1997b) is intended for use in psychiatric research settings where psychotic disorders are not expected (e.g., in a clinical trial of psychotherapy for anxious individuals), such that a simplification of the SCID-I/P's psychotic disorder module is considered sufficient. The nonpatient version of the SCID-I (SCID-I/NP; First, Spitzer, Gibbon, & Willams, 1997a) is for research participants outside of psychiatric settings (e.g., research in a primary medical care setting), with no implicit assumptions about the presence or absence of psychiatric complaints.

All versions of the SCID begin with open-ended questions regarding demographic information, work

history, chief complaint, history of present and past periods of mental illness, and assessment of current functioning. This less structured portion of the interview allows for rapport building, and it can provide helpful context for the interpretation of subsequent answers in the diagnostic section (Summerfeldt et al., 2010). The SCID comprises nine diagnostic modules: Mood Episodes, Psychotic Symptoms, Psychotic Disorders Differential, Mood Disorders Differential, Substance Use, Anxiety, Somatoform Disorders, Eating Disorders, and Adjustment Disorders.

Each diagnostic section consists of initial probes and possible follow-up questions. Every probe is explicitly linked to a diagnostic criterion item for a specific *DSM–IV* disorder. Interviewees are rated on each criterion on the basis of their answers to the probes and follow-up queries, and every criterion item may be rated as being absent, present at a subthreshold level (i.e., not at a clinically significant level), existing at or above the threshold (i.e., at a clinically significant level), or as not having sufficient evidence from the interviewee for scoring purposes. Accordingly, the SCID requires considerable clinical judgment on the part of the interviewer (Summerfeldt et al., 2010).

Zanarini et al. (2000) found excellent interrater reliability on the SCID (κs ranging from .76 to 1.0) for six *DSM–IV* diagnoses (dysthymia, any eating disorder, major depression disorder [MDD], PTSD, alcohol abuse/dependence, and drug abuse/dependence) and fair-to-good reliability (κs ranging from .57 to .65) for four diagnoses (obsessive–compulsive disorder [OCD], social phobia, generalized anxiety disorder [GAD], and panic disorder). They also found excellent test–retest reliability (κs ranging from .76 to .78) for three diagnoses (drug abuse/dependence, alcohol abuse/dependence, and PTSD), fair-to-good reliability (κ ranging from .44 to.65) for six diagnoses (GAD, social phobia, OCD, MDD, any eating disorder, and panic disorder), and poor reliability (κ = .35) for dysthymia.

The aforementioned reliability study compares somewhat favorably with an earlier multisite investigation conducted by the SCID's originators (Williams et al., 1992) utilizing diagnostic criteria from the revised third edition of the *DSM* (American Psychiatric Association, 1987). They observed only fair-to-good diagnostic reliability across Axis I diagnoses in a patient sample (κ = .61) but relatively poor reliability in a nonpatient sample (κ = .37). Test–retest reliabilities for given diagnoses ranged from a high of .86 for bulimia nervosa to a low of .40 to dysthymia.

Few published validity studies of the SCID have appeared, presumably because the SCID evinces high *face validity,* as the interview is explicitly derived from and tethered to the *DSM*-based diagnostic criteria of interest. Although available studies generally support the SCID's construct and criterion validity (see Rogers, 2001), more research is needed on this front, as is research to clarify the instrument's predictive validity across differing patient and nonpatient populations (e.g., Parks, Kmetz, & Hillard, 1995).

Diagnostic Interview Schedule (DIS)

The DIS (Robins, Helzer, Croughan, & Ratcliff, 1981) is a fully structured interviewing tool initially designed for the National Institute of Mental Health Epidemiological Catchment Area Study. The budget limitations of the study necessitated the use of laypersons to conduct all interviews (Summerfeldt et al., 2010). As a result, the researchers wanted the DIS to be as fully structured as possible, with minimal need for clinical judgment. Thus, the DIS is designed for use by both trained professionals and laypersons to assess psychiatric disorders. Although originally designed in concert with diagnostic criteria from the third edition of the *DSM*, the current iteration of the DIS (the DIS–IV; Robins, Cottler, Bucholz, & Compton, 1996) consists of 19 modules that address more than 30 distinct *DSM–IV* Axis I diagnoses as well as antisocial personality disorder on Axis II. Each module can be used independently, freeing interviewers to only look at those areas in which they are interested.

The DIS's heavily structured format has permitted the development of a computerized DIS, known as the CDIS (Robins et al., 2000). In fact, the CDIS is the only version of the DIS still in widespread use. The CDIS can be utilized either in self-administered format or as a tool for the interviewer. If the former is used, it is advised that a clinician still be present to monitor the proceedings (Kobak, Skodol, & Bender, 2008).

The DIS begins with a module assessing demographic factors. The instrument goes beyond typical demographic assessments in terms of depth, requesting information about interviewees that is not normally asked for in diagnostic interviews, and asks questions about chronological events in interviewees' past that might be connected with their current symptoms (Summerfeldt et al., 2010).

As Summerfeldt et al. (2010) have noted, because the DIS was originally developed for epidemiological studies, no chief presenting complaint is assumed. Instead, the interview proceeds through potential symptom domains in a set fashion. Likewise, all questions are asked in a specific way. If the interviewee is unclear about the meaning of any question, in keeping with the heavily structured nature of the instrument, the question is simply repeated with the exact same wording. All questions regarding the experience of the target symptom are formatted so that interviewee answers will be either "yes" or "no." Depending on the response given, follow-up probe questions may be asked to establish whether the basis for the symptom is psychiatric in nature (as opposed to being caused by a nonpsychiatric ailment or a drug) and, if so, whether the symptom occurs at a clinically significant level.

On the basis of responses provided, each symptom is coded as one of the following: not occurring; not occurring at a clinically significant level; resulting from medication, drug, or alcohol use; resulting from a physical ailment; or representing a possible psychiatric symptom. If enough symptoms for this disorder are rated as "possibly psychiatric" to warrant a potential diagnosis from the *DSM–IV*, more standardized follow-up questions are asked. Answers to all queries are eventually analyzed by the DIS's computer algorithm, which promptly delivers a diagnostic report.

Most data on DIS reliability and validity stem from pre-*DSM–IV* versions of the measure, and the relevant literature has been both mixed and controversial (Groth-Marnat, 2009; Summerfeldt et al., 2010). Reliability indices have generally fallen in the lower end of the fair-to-good range. With regard to validity, there is some evidence (Eaton, Neufeld, Chen, & Cai, 2000; Groth-Marnat, 2009; Murphy, Monson, Laird, Sobol, & Leighton, 2000) that the specificity of the DIS (its ability to detect accurately people who are *not* suffering from a given mental illness) is superior to its sensitivity (its ability to detect individuals who *are* suffering from a given mental illness). See Groth-Marnat (2009) for an in-depth discussion of these issues.

Schedules for Clinical Assessment in Neuropsychiatry (SCAN)

The SCAN (World Health Organization, 1994) is a set of instruments developed to allow dimensional ratings across an array of psychological symptoms. In principle, these ratings can be used to diagnose patients across a broad range of psychopathologies and diagnostic systems—for example, *DSM–IV* and the International Classification of Diseases (10th revision; ICD–10; World Health Organization, 2004).

The SCAN consists of four textual components and a computer scoring system. The textual components include the Present State Examination (PSE), the Item Group Checklist (IGC), a glossary of terms, and an optional Clinical History Schedule (CHS). The most substantial portion of the SCAN is the PSE, a semistructured interview. The PSE itself consists of two parts. The first part collects information on nonpsychotic symptoms (those pertaining to anxiety, mood, substance abuse, etc.). The second measures symptoms of psychosis, cognitive disorders, and disturbances in speech or behavior. All symptoms are assessed regarding the extent to which they are currently present or have existed at some point in the past. The IGC collects information from sources other than the patient (e.g., case reports), either to supplement the PSE or to replace aspects of it (albeit imperfectly) if the PSE cannot be fully completed. The glossary facilitates scoring of subject responses for each item in the SCAN. The CHS collects information on one's developmental and social history (e.g., childhood, education, intelligence), which is necessary for differentially diagnosing in many diagnostic systems (Kobak et al., 2008). Finally, the SCAN's CATEGO5 computer scoring system can provide profiles of symptoms germane to various *DSM–IV* and ICD–10 diagnostic categories.

The original field testing of the SCAN found high interrater and test–retest reliabilities (Wing, Sartorius, & Der, 1998). Regarding *DSM*-based depression and anxiety disorders, overall current and lifetime diagnoses reliabilities for SCAN have been reported as .67 and .60, respectively (Kobak et al., 2008). Andrews, Peters, Guzman, and Bird (1995) compared the SCAN with the more heavily structured Composite International Diagnostic Interview (CIDI; an interview similar to the DIS, especially in its ability to be used by nonclinicians), and observed that the CIDI evinces higher levels of interrater agreement. The less structured nature of the SCAN, however, may render it more likely to pick up on certain symptoms, particularly those pertinent to the diagnosis of mood disorders (Eaton et al. 2000).

In summarizing the validity research on the SCAN, Summerfeldt et al. (2010) stated that many consider it a "benchmark" for judging the validity of other diagnostic tests (but see Kobak et al., 2008, for a more cautionary take).

Anxiety Disorders Interview Schedule for *DSM–IV* (ADIS-IV)

The Anxiety Disorders Interview Schedule for *DSM–IV* (ADIS-IV; Brown, Di Nardo, & Barlow, 1994) provides more in-depth assessment of anxiety disorders than other notable structured and semistructured diagnostic tools. It is a semistructured, clinician-administered interview that differentially assesses for anxiety disorders as well as mood, somatoform, and substance use illnesses "because of their high comorbidity with anxiety-related diagnoses" (Summerfeldt et al., 2010, p. 101). The ADIS-IV comes in both the standard version and the Lifetime version (ADIS-IV-L; Di Nardo, Brown, & Barlow, 1994), the former of which assesses for current disorders whereas the latter assesses for both current and past problems. (Summerfeldt et al., 2010).

With respect to the assessment of *DSM-IV* diagnoses of anxiety disorders and affective illnesses, Brown, Di Nardo, Lehman, and Campbell (2001) found acceptably high levels of interrater reliability for the ADIS-IV-L, ranging from .60 to .82 (with the exception of $\kappa = .22$ for dysthymic disorder). There have appeared no published studies to date that directly address the validity of the ADIS-IV.

Schedule for Affective Disorders and Schizophrenia (SADS)

The SADS (Endicott & Spitzer, 1978) is a widely used, semistructured, clinician-administered diagnostic tool. It assesses 23 major diagnostic categories covered by the formerly influential Research Diagnostic Criteria (RDC; Spitzer, Endicott, & Robins, 1978). Although it offers diagnoses in fewer areas than many other general structured and semistructured interviews, it offers particularly extensive coverage of mood disorders. Also, just as several different versions of the SCID exist to suit the varying purposes of researchers and clinicians, the SADS comes in three main versions: (a) the standard SADS, which devotes Part I to covering mental illness that have occurred within the preceding year and Part II to those existing before that point; (b) the Lifetime version, or SADS-L, which is similar to Part II of the regular SADS, except that it also covers current problems with mental illness (albeit in less detail than Part I of the regular SADS); and (c) the Change version, or SADS-C, which focuses on temporal changes in symptom patterns.

The SADS has been popular in research circles since its inception, although its use has been limited by the relatively substantial amount of clinical expertise required of the interviewer, its lengthy administration time (2–4 hr for psychiatric patients), and its being derived from RDC as opposed to *DSM-IV* criteria. This latter concern is perhaps less worrisome for clinicians and researchers who study mood and psychotic-spectrum disorders, as the criteria for these categories in the RDC closely parallel those of the *DSM–IV*. However, particularly for diagnoses regarding anxiety and somatoform disorders, additional items are needed in the assessment to more closely connect the results of the SADS with *DSM* criteria (Summerfeldt et al., 2010).

The SADS generally demonstrates high levels of interrater (Endicott & Spitzer, 1978) and test–retest reliabilities (Spiker & Ehler, 1984). In terms of construct and content validity, the SADS has been successfully used to detect family patterns of OCD (Bienvenu et al., 2000), schizophrenia (Kendler, Gruenberg, & Kinney, 1994; Stompe, Ortwein-Swoboda, Strobl, & Friedmann, 2000), and panic disorder (Coryell, Pine, Fyer, & Klein, 2006). It has

also been observed to "predict the course, clinical features, and/or outcome in schizophrenia (Loebel et al., 1992; Stompe et al., 2000), major depression (Coryell et al., 1994), and bipolar disorder (Vieta et al., 2000; Weisman et al., 2002)" (Summerfeldt et al., 2010, p. 122).

The Clinical Interview as Therapeutic Intervention

Regardless of its degree of structure, the clinical interview allows the interviewer to begin acting as a potential agent of therapeutic change from the very first moments of the clinical interaction (Ilardi & Branstetter, 2004). In other words, the interview process has considerable implicit therapeutic promise.

Perhaps most obviously, the initial interview provides the skillful clinician with an opportunity to begin building (or strengthening) the therapeutic alliance with each patient. In fact, Carl Rogers (1961) suggested a set of principles that can help the clinician establish the alliance during any interview process, among them: maintaining a nonjudgmental attitude, viewing the patient with unconditional positive regard, reflecting accurate empathy, and conveying a sense of authenticity and genuineness. More recently, Othmer and Othmer (1994) elaborated a set of similar principles that include: putting the patient at ease, determining the source of the patient's suffering, showing appropriate empathy, assessing the patient's own understanding of his or her problems, communicating a sense of being "on their side," and acting as a credible clinical expert. Accordingly, the interview may afford the clinician an opportunity to increase the patient's sense of hope (e.g., through the perception that a caring expert is committed to understanding and helping them). In fact, in their seminal book, *Persuasion and Healing*, Frank and Frank (1991) hypothesized that increased patient hope in the early stages of therapy may be the best predictor of later benefits in therapy and mobilizes the patient to work toward change with the clinician.

Although the aforementioned therapeutic principles can apply to virtually any interview-based interaction, in recent decades clinical researchers have begun to explore the potential of specific interviewing techniques to serve as efficacious stand-alone interventions. The most notable development in this regard is that of *motivational interviewing* (MI), developed by Miller and Rollnick (2002), a set of techniques now in widespread use with individuals suffering from various substance use disorders.

Miller and Rollnick (2002) defined MI as "a client-centered, directive method for enhancing intrinsic motivation to change by exploring and resolving ambivalence" (p. 25). It is interesting enough that this clinical approach has its roots in Rogerian client-centered therapy. It focuses on the client's present interests and concerns, rather than teaching skills, changing cognitions, or discussing events of the past. However, MI is still decidedly more directive in its approach than traditional Rogerian therapy, as the clinician intentionally maneuvers to draw attention to the client's ambivalence about change and also reinforces change talk to help the client move in that direction. MI is not, however, a way of tricking people into doing something they do not want to do. Instead, it comprises a set of techniques to enhance the client's intrinsic motivation for change through various communication techniques.

As noted, the MI perspective places a central focus on the construct of ambivalence, inasmuch as clients often have decidedly mixed feelings about undertaking behavioral changes (clinician recommended or otherwise). In fact, the principal goal of MI is to help the client overcome ambivalence toward potentially salubrious change and to facilitate the development of the conditions necessary for the desired change to occur. Thus, for the MI therapist, it is crucial to serve as a catalyst for resolving a client's inherent conflict between the desire to change and the perceived costs of change (Miller & Rollnick, 2002).

Accordingly, MI makes use of four general principles to help individuals resolve such ambivalence. The first principle is expressing empathy, including a clear understanding the client's perspective of their presenting problem. The second principle involves developing the client's sense of discrepancy between how they want their life to be and any current behaviors that may be interfering with this goal. The third principle is to "roll with resistance"; that is, to consider client opposition to change a natural

process worthy of empathic validation. The fourth principle is to support the client's self-efficacy. This includes not only encouraging change talk and helping the client move toward change but also accepting the client's potential decision not to change (Miller & Rollnick, 2002).

MI is most frequently used with individuals suffering from alcohol dependence or other substance-related disorders, and the majority of clinical trials of MI's efficacy have occurred with these clinical populations. MI-based interventions have generally been found to be equivalent to other credible alternative treatments and yield a small-to-moderate effect size when compared to no-treatment conditions or placebo controls for alcohol, drug, and weight-related problems (Burke, Arkowitz, & Menchola, 2003; Lundahl, Kunz, Brownell, Tollefson, & Burke, 2010). MI has also been used as a prelude to cognitive–behavioral therapy—a brief pretreatment intervention designed to enhance motivation for complying with subsequent treatment—and it appears to be moderately effective as a means of increasing adherence to potentially demanding cognitive–behavioral therapy protocols (Burke, Arkowitz & Dunn, 2002). Likewise, MI has been shown to improve adherence to a variety of medical treatments, including those for managing diabetes (Resnicow et al., 2002). However, MI interventions have not shown to be particularly effective for smoking or HIV-risk behaviors (Burke et al., 2003; Lundahl et al., 2010).

It remains for future researchers to clarify the extent to which MI-based interventions may be useful in the treatment of other clinical populations (e.g., those with eating disorders) and to elucidate the salient mediators of MI-derived therapeutic effects. Likewise, further investigation is warranted to help identify the extent to which other interview-based clinical techniques—beyond those developed within the MI framework—may be useful as therapeutic interventions in their own right.

CONCLUSION

The clinical interview is a form of conversation with an explicit therapeutic purpose. It has been regarded as a foundational element of psychological and psychiatric practice for over a century, integral to the core processes of both assessment and intervention. However, there is no single monolithic structure or content that characterizes all clinical interviews; on the contrary, researchers have identified three distinctive categories into which interviews are now commonly subclassified: *structured*, *unstructured*, and *therapeutic*.

The unstructured interview is only minimally directed by the clinician. It evinces an open, free-flowing, conversational dynamic between patient and clinician. The unstructured format is of ancient provenance—finding echoes in the medical interviews of clinicians such as Pinel and Rush from centuries past. Perhaps it is not surprising that it was also the dominant mode of interviewing across much of the 20th century in mental health settings. Although the unstructured format allows for a high degree of clinician creativity, flexibility, and spontaneity, it has been criticized on two important grounds: (a) It is among the least reliable and valid of extant clinical assessment procedures, and (b) it appears to require a lengthy training regimen to master the process. Nevertheless, the unstructured format still enjoys widespread use among practicing clinicians.

Structured interviewing, on the other hand, is characterized by a predetermined set of queries that the clinician is directed to ask verbatim, in a precisely defined order. Such interviews have become much more widely utilized in both applied and research settings in recent decades, in part because of their superior psychometric properties. Commonly used structured interviews include the SCID, the DIS, the SCAN, the ADIS-IV, and the SADS.

There also exist a number of clinical interview techniques that serve as stand-alone modes of therapeutic intervention. MI describes the most widely used among such interview-as-intervention strategies. MI encompasses an array of interview techniques designed to enhance the patient's intrinsic motivation for salubrious change; for example, through drawing attention to the patient's likely ambivalence about the behavioral change process itself.

Although MI has received some measure of empirical support regarding its efficacy, it remains for future investigators to demonstrate the therapeutic potential—or lack thereof—of the many non-MI

interview-based techniques. Likewise, unstructured and structured interviewing both remain fertile domains of active research investigation.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Andrews, G., Peters, L., Guzman, A., & Bird, K. (1995). A comparison of two structured diagnostic interviews: CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry, 29*, 124–132. doi:10.3109/00048679509075901

Bagby, R. M., Wild, N., & Turner, A. (2003). Psychological assessment in adult mental health settings. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, pp. 213–234). Hoboken, NJ: Wiley.

Basco, M. R. (2003). Is there a place for research diagnostic methods in clinic settings? In J. M. Oldham & M. B. Riba (Eds.), *Review of psychiatry* (Vol. 22, pp. 1–28). Washington, DC: American Psychiatric Press.

Beutler, L. E. (1995). The clinical interview. In L. E. Beutler & M. R. Berren (Eds.), *Integrative assessment of adult personality* (pp. 94–120). New York, NY: Guilford Press.

Bienvenu, O. J., Samuels, J. F., Riddle, M. A., Hoehn-Saric, R., Liang, K., Cullen, B. A. M., & Nestadt, G. (2000). The relationship of obsessive–compulsive disorder to possible spectrum disorders: Results from a family study. *Biological Psychiatry, 48*, 287–293. doi:10.1016/S0006-3223(00)00831-3

Brown, T. A., Di Nardo, P. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview Schedule for DSM–IV (ADIS-IV)*. New York, NY: Oxford University Press.

Brown, T. A., Di Nardo, P. A., Lehman, C. L., & Campbell, L. A. (2001). Reliability of *DSM–IV* anxiety and mood disorders: Implications for the classification of emotional disorders. *Journal of Abnormal Psychology, 110*, 49–58. doi:10.1037/0021-843X.110.1.49

Burke, B. L., Arkowitz, H., & Dunn, C. (2002). The efficacy of motivational interviewing and its adaptations. In W. R. Miller & S. Rollnick (Eds.), *Motivational interviewing: Preparing people for change* (2nd ed., pp. 217–250). New York, NY: Guilford Press.

Burke, B. L., Arkowitz, H., & Menchola, M. (2003). The efficacy of motivational interviewing: A meta-analysis of controlled clinical trials. *Journal of Consulting and Clinical Psychology, 71*, 843–861. doi:10.1037/0022-006X.71.5.843

Carlat, D. J. (2005). *The psychiatric interview: A practical guide* (2nd ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.

Coryell, W., Pine, D., Fyer, A., & Klein, D. (2006). Anxiety responses to CO_2 inhalation in subjects at high-risk for panic disorder. *Journal of Affective Disorders, 92*, 63–70. doi:10.1016/j.jad.2005.12.045

Coryell, W., Winokur, G., Maser, J. D., Akiskal, H. S., Keller, M. B., & Endicott, J. (1994). Recurrently situational (reactive) depression: A study of course, phenomenology and familial psychopathology. *Journal of Affective Disorders, 31*, 203–210. doi:10.1016/0165-0327(94)90030-2

Craig, R. J. (2003). Assessing personality and psychopathology with interviews. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, pp. 487–508). Hoboken, NJ: Wiley.

Di Nardo, P. A., Brown, T. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview Schedule for DSM–IV: Lifetime version.* New York, NY: Oxford University Press.

Eaton, W. W., Neufeld, K., Chen, L., & Cai, G. (2000). A comparison of self-report and clinical diagnostic interviews for depression. *Archives of General Psychiatry, 57*, 217–222. doi:10.1001/archpsyc.57.3.217

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 35*, 837–844. doi:10.1001/archpsyc.1978.01770310043002

Ferriter, M. (1993). Computer aided interviewing and the psychiatric social history. *Social Work and Social Sciences Review, 4*, 255–263.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1996). *Structured Clinical Interview for DSM–IV Axis I Disorders, Clinical Version (SCID-CV)*. Washington, DC: American Psychiatric Press.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997a). *Structured Clinical Interview for DSM–IV Axis I Disorders, Research Version, Nonpatient Edition (SCID-I/NP)*. New York, NY: Biometrics Research, New York Psychiatric Institute.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997b). *Structured Clinical Interview for DSM–IV Axis I Disorders, Research Version, Patient Edition With Psychotic Screen (SCID-I/P W/Psy Screen)*. New York, NY: Biometrics Research, New York Psychiatric Institute.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997c). *Structured Clinical Interview for DSM–IV Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. New York, NY: Biometrics Research, New York Psychiatric Institute.

First, M. B., Spitzer, R. L., Williams, J. B. W., & Gibbon, M. (1995). *Structured Clinical Interview for DSM–IV, Research Version (SCID-I)*. Washington, DC: American Psychiatric Press.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). Hoboken, NJ: Wiley.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). New York, NY: Wiley.

Helzer, J. E., Robins, L. N., Taibleson, N., Woodruff, R., A., Reich, T., & Wish, E. D. (1977). Reliability in psychiatric diagnosis. *Archives of General Psychiatry, 34*, 129–133. doi:10.1001/archpsyc.1977.01770140019001

Hersen, M., & Bellack, A. S. (1988). DSM-III and behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 67–84). Needham Heights, MA: Allyn & Bacon.

Hong, P. Y., & Ilardi, S. S. (2004). Structured and semistructured clinical interviews. In W. E. Craighead & C. B. Nemeroff (Eds.), *Concise Corsini encyclopedia of psychology and behavioral science* (pp. 954–957). New York, NY: Wiley.

Huffcutt, A. I., & Arthur, W. (1994). Arthur & Arthur (1984). revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184–190. doi:10.1037/0021-9010.79.2.184

Ilardi, S. S., & Branstetter, A. D. (2004). Unstructured clinical interview. In E. Craighead & C. B. Nemeroff (Eds.), *Concise Corsini encyclopedia of psychology and behavioral science* (pp. 1016–1019). Hoboken, NJ: Wiley.

Jenkins, S. (2007). Unstructured interviewing. In M. Hersen & J. C. Thomas (Eds.), *Handbook of clinical interviewing with adults* (pp. 7–23). Thousand Oaks, CA: Sage. doi:10.4135/9781412982733.n2

Johnson, W. R. (1981). Basic interviewing skills. In C. E. Walker (Ed.), *Clinical practice of psychology: A guide for mental health professionals* (pp. 83–128). Elmsford, NY: Pergamon Press.

Jones, K. D. (2010). The unstructured clinical interview. *Journal of Counseling and Development, 88*, 220–226. doi:10.1002/j.1556-6678.2010.tb00013.x

Kanfer, F. H., & Scheft, B. K. (1988). *Guiding the process of therapeutic change.* Champaign, IL: Research Press.

Kendler, K. S., Gruenberg, A. M., & Kinney, D. (1994). Independent diagnoses of adoptees and relatives as defined by DSM-III in the provincial and national samples of the Danish Adoption Study of Schizophrenia. *Archives of General Psychiatry, 51*, 456–468. doi:10.1001/archpsyc.1994.03950060020002

Kobak, K. A., Skodol, A. E., & Bender, D. S. (2008). Diagnostic measures for adults. In A. J. Rush Jr., M. B. First, & D. Blacker (Eds.), *Handbook of psychiatric measures* (2nd ed., pp. 35–60). Washington, DC: American Psychiatric Press.

Lazarus, A. A. (1973). Multimodal behavior therapy: Treating the "BASIC ID." *Journal of Nervous and Mental Disease, 156*, 404–411. doi:10.1097/00005053-197306000-00005

Loebel, A. D., Lieberman, J. A., Alvir, J. M., Mayerhoff, D. I., Geisler, S. H., & Syzmanski, S. R. (1992). Duration of psychosis and outcome in first-episode schizophrenia. *American Journal of Psychiatry, 149*, 1183–1188.

Lundahl, B. W., Kunz, C., Brownell, C., Tollefson, D., & Burke, B. (2010). A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on Social Work Practice, 20*, 137–160. doi:10.1177/1049731509347850

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment, 1*, 18–26. doi:10.1111/j.1468-2389.1993.tb00080.x

Maruish, M. E. (2008). The clinical interview. In R. P. Archer & S. R. Smith (Eds.), *Personality assessment* (pp. 37–80). New York, NY: Routledge Press.

Matarazzo, J. D. (1965). The interview. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 403–450). New York, NY: McGraw-Hill.

Miller, C. (2003). Interviewing strategies. In M. Hersen & S. M. Turner (Eds.), *Diagnostic interviewing* (3rd ed., pp. 47–66). New York, NY: Kluwer Academic.

Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people for change* (2nd ed.). New York, NY: Guilford Press.

Mohr, D., & Beutler, L. E. (2003). The integrative clinical interview. In L. E. Beutler & G. Groth-Marnat (Eds.), *Integrative assessment of adult personality* (2nd ed., pp. 82–122). New York, NY: Guilford Press.

Murphy, J. M., Monson, R. R., Laird, N. M., Sobol, A. M., & Leighton, A. H. (2000). A comparison of diagnostic interviews for depression in the Stirling County study: Challenges for psychiatric epidemiology. *Archives of General Psychiatry, 57*, 230–236. doi:10.1001/archpsyc.57.3.230

Nuttall, E., & Ivey, A. E. (1986). The diagnostic interview process. In H. M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 105–140). New York, NY: Guilford Press.

O'Brien, W. H., & Tabaczynski, T. (2007). Unstructured interviewing. In M. Hersen & J. C. Thomas (Eds.), *Handbook of clinical interviewing with children* (pp. 16–29). Thousand Oaks, CA: Sage.

Othmer, E., & Othmer, S. C. (1994). *The clinical interview using DSM-IV: Vol. 1. Fundamentals*. Washington, DC: American Psychiatric Press.

Parks, J. J., Kmetz, G., & Hillard, J. R. (1995). Underdiagnosis using SCID-R in the homeless mentally ill. *Psychiatric Quarterly, 66*, 1–8. doi:10.1007/BF02238712

Porter, E. H. (1950). *An introduction to therapeutic counseling*. Boston, MA: Houghton-Mifflin.

Resnicow, K., DiIorio, C., Soet, J. E., Borrelli, B., Ernst, D., Hecht, J., & Thevos, A. K. (2002). Motivational interviewing in medical and public health settings. In W. R. Miller & S. Rollnick (Eds.), *Motivational interviewing: Preparing people for change* (2nd ed., pp. 251–269). New York, NY: Guilford Press.

Robins, L. N., Cottler, L. B., Bucholz, K. K., & Compton, W. (1996). *The diagnostic interview schedule (Version IV)*. St. Louis, MO: Washington University School of Medicine.

Robins, L. N., Cottler, L. B., Bucholz, K. K., Compton, W., North, C., & Rourke, K. (2000). *The diagnostic interview schedule for DSM-IV (DIS-IV)*. St. Louis, MO: Washington University School of Medicine.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381–389. doi:10.1001/archpsyc.1981.01780290015001

Rogers, C. (1961). *On becoming a person*. Boston, MA: Houghton-Mifflin.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford Press.

Rosqvist, J., Bjorgvinsson, T., & Davidson, J. (2007). Philosophical underpinnings of clinical interviewing. In M. Hersen & J. C. Thomas (Eds.), *Handbook of clinical interviewing with adults* (pp. 2–6). Thousand Oaks, CA: Sage. doi:10.4135/9781412982733.n1

Ruggeri, M., Dall'Agnola, R., Agostini, C., & Bisoffi, G. (1994). Acceptability, sensitivity and content validity of the VECS and VSSS in measuring expectations and satisfaction in psychiatric patients and their relatives. *Social Psychiatry and Psychiatric Epidemiology, 29*, 265–276. doi:10.1007/BF00802049

Segal, D. L. (1997). Structured interviewing and *DSM* classification. In S. M. Turner & M. Hersen (Eds.), *Adult psychopathology and diagnosis* (3rd ed., pp. 24–57). New York, NY: Wiley.

Segal, D. L., & Falk, S. B. (1998). Structured interviews and rating scales. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 158–178). Needham Heights, MA: Allyn & Bacon.

Segal, D. L., Hersen, M., & Van Hasselt, V. B. (1994). Reliability of the structured clinical interview for *DSM-III-R*: An evaluative review. *Comprehensive Psychiatry, 35*, 316–327. doi:10.1016/0010-440X(94)90025-6

Shea, S. C. (1988). *Psychiatric interviewing: The art of understanding*. Philadelphia, PA: W. B. Saunders.

Shea, S. C. (1990). Contemporary psychiatric interviewing: Integration of DSM-III-R, psychodynamic concerns, and mental status. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed., pp. 283–307). New York, NY: Pergamon Press.

Shedler, J. (2002). A new language for psychoanalytic diagnosis. *Journal of the American Psychoanalytic Association, 50*, 429–456. doi:10.1177/00030651020500022201

Snyder, W. V. (1945). An investigation of the nature of nondirective psychotherapy. *Journal of General Psychology, 33*, 139–223.

Sommers-Flanagan, J., & Sommers-Flanagan, R. (2003). *Clinical interviewing* (3rd ed.). Hoboken, NJ: Wiley.

Spiker, D. G., & Ehler, J. G. (1984). Structured psychiatric interviews for adults. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 291–304). New York, NY: Pergamon Press.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria. *Archives of General Psychiatry, 35*, 773–782. doi:10.1001/archpsyc.1978.01770300115013

Stompe, T., Ortwein-Swoboda, G., Strobl, R., & Friedmann, A. (2000). The age of onset of schizophrenia and the theory of anticipation. *Psychiatry Research, 93*, 125–134. doi:10.1016/S0165-1781(00)00103-7

Sullivan, H. S. (1954). *The psychiatric interview*. New York, NY: Norton.

Summerfeldt, L. J., Kloosterman, P. H., & Antony, M. M. (2010). Structured and semistructured diagnostic interviews. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (2nd ed., pp. 95–137). New York, NY: Guilford Press.

Van Dyke, D. (2005). The family therapy interview. In R. J. Craig (Ed.), *Clinical and diagnostic interviewing* (2nd ed., pp. 131–142). Lanham, MD: Jason Aronson.

Vieta, E., Colom, F., Martinez-Aran, A., Benabarre, A., Reinares, M., & Gasto, C. (2000). Bipolar II disorder and comorbidity. *Comprehensive Psychiatry, 41*, 339–343. doi:10.1053/comp.2000.9011

Vitiello, B., Malone, R., Buschle, P. R., & Delaney, M. A. (1990). Reliability of DSM-III diagnoses of hospitalized children. *Hospital and Community Psychiatry, 41*, 63–67.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1966). *Pragmatics of human communication*. New York, NY: Norton.

Weisman, A., Tompson, M. C., Okazaki, S., Gregory, J., Goldstein, M. J., Rea, M., & Miklowitz, D. J. (2002). Clinicians' fidelity to a manual-based family treatment as a predictor of the one-year course of bipolar disorder. *Family Process*, *41*, 123–131.

Whitaker, R. (2010). *Mad in America: Bad science, bad medicine, and the enduring mistreatment of the mentally ill*. New York, NY: Basic Books.

Widiger, T. A. (2008). Clinical interviews. In A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 47–65). New York, NY: Oxford University Press.

Widiger, T. A., Sanderson, C., & Warner, L. (1986). The MMPI, prototypal typology, and borderline personality disorder. *Journal of Personality Assessment, 50*, 540–553. doi:10.1207/s15327752jpa5004_2

Wiens, A. (1976). The assessment interview. In I. B. Weiner (Ed.), *Clinical methods in psychology* (pp. 3–60). Oxford, England: Wiley.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., . . . Wittchen, H. (1992). The Structured Clinical Interview for DSM-III-R: Multisite test-retest reliability. *Archives of General Psychiatry*, *49*, 630–636. doi:10.1001/archpsyc.1992.01820080038006

Wing, J. K., Sartorius, N., & Der, G. (1998). International field trials: SCAN-O. In J. K. Wing, N. Sartorius, & T. B. Üstün (Eds.), *Diagnosis and clinical measurement in psychiatry: A reference manual for SCAN* (pp. 86–109). Cambridge, England: Cambridge University Press.

Wolpe, J. (1973). *The practice of behavior therapy*. New York, NY: Pergamon Press.

World Health Organization. (1994). *Schedules for clinical assessment in neuropsychiatry: Version 2.0 manual*. Geneva, Switzerland: Author.

World Health Organization. (2004). *International classification of diseases* (10th revision). Geneva, Switzerland: Author.

Young, J. G., O'Brien, J. D., Gutterman, E. M., & Cohen, P. (1987). Research on the clinical interview. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 613–620. doi:10.1097/00004583-198709000-00002

Zanarini, M. C., Skodol, A. E., Bender, D., Dolan, R., Sanislow, C., Schaefer, E., . . . Gunderson, J. G. (2000). The Collaborative Longitudinal Personality Disorders Study: Reliability of Axis I and Axis II diagnoses. *Journal of Personality Disorders*, *14*, 291–299. doi:10.1521/pedi.2000.14.4.291

CHAPTER 8

ASSESSMENT OF INTELLECTUAL FUNCTIONING IN ADULTS

Phillip L. Ackerman

At the outset, it must be noted that the intelligence of adults is fundamentally different in many ways from the intelligence of children and adolescents. There are several reasons for this distinction, and these are discussed in this chapter. The first part of the chapter presents a review of the modern development of tests of intelligence for children and how such tests were initially adapted for the assessment of adults. Next discussed are some of the early findings about adult intelligence, and how these related to later developments of theory and measures specifically designed to assess adult intelligence. In addition, major sources of modern empirical research and theory that inform the interpretations of adult intelligence measures are reviewed. In a final section, some enduring challenges associated with assessing intelligence in adults are described.

HISTORICAL BACKGROUND

Modern assessment of intelligence started with the seminal work by Binet and his colleagues (most notably, T. Simon) in the early 1900s. Binet was given the task of developing assessments of children to determine which students were unlikely to benefit from mainstream classroom instruction in the Paris schools. Before Binet's development of intelligence scales, diagnoses of mental retardation were made on the basis of a "medical" method, "which aims to appreciate the anatomical, physiological, and pathological signs of inferior intelligence" (Binet & Simon, 1905/1973, p. 40). Binet observed that this particular approach was not altogether scientific or valid, and he set out to develop a set of scales that were valid for prediction of academic success. However, Binet noted that there were two other approaches to assessing intelligence that one could make—one he called the "pedagogical method" and the other he called the "psychological method." The pedagogical method of assessing a child's intelligence "aims to judge of the intelligence according to the sum of acquired knowledge" (p. 40)—that is, finding out what the child *knows*—the corpus of the child's knowledge and skills (e.g., reading, writing, problem solving). The psychological method, according to Binet, "makes direct observations and measurements of the degree of intelligence" (p. 40). For Binet, this meant assessing individual differences in memory, reasoning, common cultural knowledge, verbal comprehension and production, and so on.

Binet ultimately selected the psychological method of assessing intelligence over the pedagogical method partly because he wanted to eliminate, at least as far as possible, the influences of socioeconomic status on the measurement of intelligence. These influences would have been strongly associated with literacy, for example, among young Paris school children at the turn of the last century.

Two major criteria were selected by Binet for the validation of the intelligence scales he and Simon created. The first is called "age differentiation." That is, Binet's fundamental assumption was that intellectual capabilities in children increased with age on average, so that older children were expected to perform better than younger children on the same

DOI: 10.1037/14048-008
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

items. Items created for his intelligence scale were selected if they met this criterion and rejected if older children did not perform better than younger children. Scales of intelligence were also retained if they were associated with success in school—that is, the test scales were selected if they were correlated with the criterion of success/failure in school. Binet indicated that scores on his intelligence scales were provided in terms of "mental age"—that is, an individual's score was based on the average level of performance obtained by a group of children of various different ages. A student with a mental age of 10 performed as well as the average 10-year-old in the normative sample.

The Binet–Simon scales were immensely successful in predicting academic success/failure in children and adolescents. Within 10 to 15 years of their creation, for example, the scales had been translated into English and refined to result in measures still in wide use today (e.g., the Stanford–Binet measure, first published by in Terman, 1916, and most recently revised in 2003; see Roid, 2003). These tests show a clear lineage in both their underlying fundamental design and application to those designed by Binet and Simon in 1905.

As discussed in substantial detail in this chapter, the choices of the psychological method, scores of mental age, and the criteria of age differentiation and school achievement in the development of intelligence measures for children had major, but not necessarily positive, influences on the assessment of intelligence in adults for much of the past century.

EARLY MEASURES OF ADULT INTELLIGENCE

Although several researchers explored the extension of the Binet-type scales to older adolescents and adults in the 1910s, the first major assessment of adult intelligence was undertaken by the U.S. Army during World War I, using primarily the Army Alpha test. The test generally mirrored the Binet scales in content (e.g., tests of arithmetic, analogies, general information, synonyms and antonyms), but two major modifications were included. First, the test items were written in a way that made the content more suitable to adults than to children: For example, "If you buy two packages of tobacco at 7 cents each and a pipe for 60 cents, how much change should you get from a two-dollar bill?" (Yoakum & Yerkes, 1920, p. 206) and "The Knight engine is used in the Packard . . . Lozier . . . Stearns . . . Pierce Arrow" (Yoakum & Yerkes, 1920, p. 219). Second, in contrast to the Binet scales, which required individual administration with one examiner and one child, the Army Alpha test was administered in large group-testing environments. This change in administration meant that (a) the test required that examinees be able to read and write, and (b) many questions were provided in a multiple-choice format, so that examinees needed only to recognize the correct answer rather than produce it (e.g., see discussion by Carroll, 1982).

In addition to the Alpha test, the Army also administered the Beta test to examinees who were either "illiterate or unable to understand English" (Yoakum & Yerkes, 1920, p. 51) or who performed poorly on the Alpha test. The Beta test was ostensibly a nonverbal intelligence test and is largely composed of tests developed earlier by investigators who were concerned with testing recent immigrants or children with hearing impairments (e.g., see Pintner & Paterson, 1917). Results from the Beta test administrations were not particularly informative, for a number of reasons. Perhaps the most important difficulty was that these tests, like the Alpha, were administered in large group settings. Instructions for the Beta test were provided by pantomime and spoken English, rendering the test situation probably quite confusing to many examinees.

Initial reports of the results from testing 1,700,000 men with the Army Alpha test were startling and somewhat controversial (e.g. see Brigham, 1922; Lippmann, 1922). When the scores on the test were compared to Stanford–Binet scores, it appeared that the average mental age of the adult Army conscripts was about 13 years—that is, the average adult male had an intellectual level that was equivalent to the average 13-year-old adolescent. Several competing interpretations were offered for these results. One interpretation was that the average adult male was not particularly bright. Another interpretation was that, as adults increase in age, intelligence scores decline. (This interpretation was

difficult to justify on the basis of the existing data, because very few of the Army conscripts were older than 50 years of age.) Ultimately, however, the consensus conclusion was that the "mental age" concept was not a useful index for intelligence after adolescence, if intelligence does not show substantial continued growth with age after adulthood is reached.

EARLY STUDIES OF ADULT INTELLIGENCE

The first major investigation of the relationship between adult ages and intelligence was undertaken by Conrad, Jones, and Hsaio (e.g., see Conrad, 1930; Hsaio, 1927; Jones & Conrad, 1933/1998). They tested a cross-section of over 1,000 people from age 10 to age 60 with the Army Alpha test. Instead of only focusing on overall scores on the test (or Binet-equivalent mental ages), these investigators also looked at raw scores on each of the eight subtests of the Army Alpha. They found that overall scores on the Army Alpha reached a peak around age 20, with lower overall scores among older examinees. In addition, they found that two subtests of the Army Alpha showed a different age-related trend. For the Synonym/Antonym test and the General Information test, declines with increasing age in adulthood were *not* observed. Thus, although older examinees performed more poorly on most of the subtests of the Army Alpha, they performed just as well on tests that depended mostly on verbal abilities and common cultural knowledge, when compared with younger adults and adolescents.

How one interprets these data results in a fundamental divergence of approaches to adult intelligence. On the one hand, Jones and Conrad (1933/1998) suggested that intelligence levels rise rapidly in childhood and adolescence and decline rapidly during adulthood. With respect to the two Army Alpha subtests that showed adults maintaining performance with increasing age, they concluded that those two "present an unfair advantage to those in the upper age brackets" (Jones & Conrad, 1933/1998, p. 170). On the other hand, one might conclude that some aspects of intelligence decline with increasing age in adulthood, but other aspects of intelligence are well preserved. Of course, one might otherwise argue that none of the Army Alpha subtests adequately represent adult intelligence, rendering the entire data set largely meaningless.

It is important to note that there is a fundamental interpretation difficulty with cross-sectional studies of this type, where older adults represent different cohorts—that is, the 50-year-old examinees in the sample were born and raised in an educational and social environment quite different from that of the 10-year-old examinees in the sample (e.g., prominent differences included access to education and mass media). Cross-sectional comparisons across age groups are thus problematic, because it is impossible to separate aging effects from age cohort differences—such influences are "confounded" (e.g., see Schaie & Strother, 1968).

From the 1920s through the 1930s, the conventional wisdom regarding intelligence was that it peaked in the mid-teenage years and declined rapidly with increasing age. For most intents and purposes, the Army Alpha test and similar instruments were the measures used to assess adult intelligence, although mostly for educational selection (i.e., by undergraduate institutions; see Ackerman, 1996, for a review), and occupational selection purposes (i.e., for job selection; see Kanfer, Ackerman, Murtha, & Goff, 1995, for a review). Other than the work by Conrad and his colleagues, few investigations were made of adult intelligence.

In 1939, however, Wechsler introduced an individual test of intelligence specifically designed for administration to adults, called the Bellevue Test (Wechsler, 1939/1944). The test was designed for myriad uses, including clinical assessments. Several characteristics of this test were divergent from those of the Binet–Simon tests and the Army Alpha test. First, Wechsler designed the test so that the content was more appropriate to older adults rather than to children and adolescents. Like the Binet–Simon tests (but not the Army Alpha), the administration required a one-to-one format and did not require the examinee to be able to read or write. Wechsler also abandoned the much-maligned mental age concept for scores on the test. Instead, he developed a measure that was norm-referenced within the adult population. An intelligence quotient (IQ) of 100 was

designated as the mean performance of adults at the same age group, and the scores were scaled to have a standard deviation of 15 points, so that an IQ of 115 had no reference to mental age but instead represented an individual whose performance placed him or her 1 *SD* above the mean for his or her age group, which translates to a percentile rank of 84. Such an individual performed better than 84% of his/her age cohort group members. Wechsler originally provided separate age norms for adolescents between 15 and 19 and for adults aged 20 to 39 and 40 to 59.

Validation of the Wechsler scales was not obtained by comparison with school grades as in the case of the Binet tests but by comparisons with scores on other tests of mental abilities, and, most notably, with case studies of clinical diagnoses, such as "organic brain diseases." For Wechsler, these assessments of adult intelligence could be used as a key piece of information to aid in the diagnoses of particular kinds of psychological and neurological pathology.

Finally, Wechsler also noted that for adults, an overall IQ score might not be as meaningful or useful (e.g., for diagnostic purposes) as separate scores for verbal and nonverbal (called "Performance") components of the intelligence test. In light of the data presented by Conrad and Jones as well as his own test results, Wechsler noted that some aspects of intellectual abilities are much better preserved with increasing adult ages than others. Abilities that involve verbal comprehension, vocabulary, general information, and so on appeared to be much better preserved with age than abilities such as short-term memory and arithmetic, perceptual speed, and spatial processing. To account for these differences, Wechsler's intelligence test provides two separate indexes, a Verbal scale and a Performance scale. In the normal population, these two scores are substantially positively correlated and can be combined to yield an overall IQ score. However, there is an advantage to examining these two scores separately, when the assessment is conducted for diagnostic purposes.

Wechsler's approach to assessment of adult intelligence revolutionized the field for clinical assessments of adults. His scales eliminated several of the key problems in adapting the Binet-type scales for adults; namely, eliminating the child and adolescent orientation of the item content, eliminating the dependence on the mental age concept, taking account of the fact that some intellectual abilities are better preserved during adulthood than others, and seeking validation beyond the school classroom. The main shortcoming of Wechsler's approach was that he did not take account of cohort differences among the adults sampled for his norms. That is, even with the revisions to content and procedures, in Wechsler's view, adult intelligence peaked in the early 20s and still declined substantially with increasing age in adulthood. A later study by Owens (1953)—the first published longitudinal study of adult intelligence—provided evidence to suggest that, for individuals, there were not steep declines in intelligence with age, at least for a significant portion of adult life. Owens administered the Army Alpha test to a group of men who had completed the same test 31 years before (when they completed the test as part of their entry to college). For this sample, average intelligence actually increased, from when the men were 19 years of age to when they were 50 years of age. The two tests identified by Hsaio (1927) as showing the smallest declines in the cross-sectional study of age and intelligence (Information and Synonym/Antonym) showed the largest increases in scores over the 31-year lag from initial test to retest on the Army Alpha test.

THEORIES OF ADULT INTELLIGENCE

Contemporaneous to, but independent of, Wechsler's development of the adult intelligence test, two related theoretical approaches to understanding the nature and progression of adult intelligence were developed. One approach, that of Hebb, was based on data from neurological assessments in clinical populations; the other approach by Cattell, was based on examination of test scores in normal populations. Both of these approaches are discussed here in turn.

Hebb

From examining patients who had experienced removal of brain tumors or other excisions of brain tissue, the neuropsychologist D. O. Hebb (1939, 1942) noted that different aspects of intellectual

functioning appeared to be most or least affected by the loss of neural tissue or associated with other neurological incidents. Hebb described these different types of adult intellectual functioning as Intelligence A and Intelligence B. For Hebb, Intelligence A was "direct intellectual power" (1942, p. 289)—the aspect of intelligence that is involved in abstract reasoning, learning new material, and similar kinds of tasks. In contrast, the characteristics of Intelligence B involve "the establishment of routine modes of response to common problems" (1942, p. 289). That is, Intelligence B represents the stored knowledge and skills that an individual has acquired over his or her lifetime, whereas Intelligence A is most highly associated with novel tasks and new learning. Hebb noted that Intelligence A is most likely to show declines with increasing age in adults, but that Intelligence B is well preserved throughout much of adult life. Moreover, Intelligence A is most likely to be impaired by neurological incidents, and Intelligence B is most likely to be more robust, in the face of neurological incidents.

To Hebb, the conclusions of Jones and Conrad regarding the unsuitability of intelligence tests that showed preserved abilities in adulthood was entirely wrong, because Intelligence B is an important determinant of what tasks an adult can accomplish. Indeed, as Hebb noted, many adults are able to function reasonably well on the basis of preserved Intelligence B, even with significant impairments in Intelligence A. To a substantial degree, Intelligence A is made up of the kinds of abilities that Wechsler assessed with his Performance Scale, and Intelligence B is made up of the kinds of abilities that Wechsler assessed with his Verbal Scale. The theoretical perspective offered by Hebb provided a firm scientific foundation for the more empirically based approach that served as a basis for Wechsler's separation of different kinds of intelligence scales.

Cattell

Almost simultaneously with Hebb's proposed Intelligence A and Intelligence B, Cattell (1943) proposed that there are two different kinds of intelligence: called *fluid intelligence* (Gf) and *crystallized intelligence* (Gc). Gf, according to Cattell, is physiologically based, develops rapidly during childhood, and declines in adulthood. Elements of Gf include the same kinds of abstract reasoning and novel learning components described by Hebb as Intelligence A. Gc is made up of the range of knowledge and skills the individual possessed and is well preserved during much of adult life, similar to Hebb's Intelligence B. Cattell further proposed that Gc develops through the actions of Gf; that is, Gf is necessary for the acquisition of knowledge that becomes Gc. In this way, Cattell provides an explanation for why it is that older adults typically do not show rapid or substantial growth in Gc, because by the time adulthood is reached, Gf is already in decline, compared with its levels during childhood and adolescence.

Although Hebb did not develop his theory of two intelligences much in subsequent years, Cattell refined and greatly expanded his theory of adult intelligence in the decades that followed. For example, he included the roles of personality and motivation in the context of child and adult intellectual development. Some additional discussion of more recent developments in this theory is presented in a later section of this chapter.

IMPLICATIONS OF THEORY FOR ADULT INTELLIGENCE ASSESSMENT

The assessment approach of Wechsler and the common elements in the Hebb and Cattell theories set the stage for a reconsideration of the consensus view that adulthood is characterized by an inevitable decline of intellectual functioning with increasing age. In addition, the longitudinal data reported by Owens suggested that the traditional means of drawing conclusions about individual aging from cross-sectional studies may substantially overstate the case for such inevitable decline of intelligence during adulthood.

Once the issues of mental age in scaling intelligence were resolved by Wechsler, it was possible to view adult intelligence as normative, in comparison to adults of similar ages, rather than in comparison with child and adolescent intelligence. The division of intelligence into two broadly differentiable kinds of abilities (i.e., Performance, Intelligence A, or Gf; Verbal, Intelligence B, or Gc) provided a new perspective on the nature of age-related differences in

intellectual functioning. The remaining difficulty with such perspectives, however, related to the contrast between Hebb's positive view on the one hand and the Jones and Conrad negative view on the other hand, regarding the suitability of tests of intelligence for adults that did not show declines in performance with increasing age.

Fundamentally, the question that lay unresolved was what "adult intelligence" should really mean. Where Binet had school failure as the ultimate criterion for intelligence assessments, and industrial–organizational psychologists (e.g., see Kanfer et al., 1995, for a review) had job performance as the criterion for intelligence assessments, no real consensus existed for whether such notions were sufficiently applicable for the construct of adult intelligence. In addition, later refinements of Cattell's theory of adult intelligence suggested that there is an underlying gap between how Gc is typically assessed and the overarching nature of Gc. Specifically, Cattell suggested that if Gc represents the entire corpus of an individual's acquired knowledge and skills, to assess individual differences in Gc, one must choose between creating a test specific to each domain of knowledge and skills that an individual might possess and using a test that only focuses on core cultural knowledge to which most individuals could be expected to have been previously exposed.

Cattell (1957) noted the infeasibility of creating innumerable tests of different aspects of Gc that would, for example, entail creating a chemistry test to assess a chemist's knowledge about chemistry, a plumbing test to assess a plumber's knowledge, and tests of domains of art, music, current events, cooking, and so on. Indeed, because different individuals have varying levels of knowledge of topics outside of their own occupational or avocational (hobby) activities, the number of different kinds of tests needed to generate a single Gc profile for an individual would clearly be beyond the capabilities of any examiner. The alternative approach for assessing Gc is to focus on what an individual knows that is, more or less, common to the culture within which the assessment is used. In fact, Wechsler used this strategy to assess abilities in the areas of vocabulary, general information, and so on. Cattell referred to this construct as "historical" Gc, to be distinguished from the theoretical but unmeasured "current" Gc. As an aside, this is the same strategy used mostly in educational selection tests for undergraduate schools and graduate schools. The SAT and the Graduate Record Examinations (GRE) both assess quantitative abilities, for example, but the content of these tests only involves algebra and geometry, although many high school seniors and, most certainly, many college students go on to acquire knowledge and skills in calculus and other advanced mathematics. These tests focus on historical Gc, in much the same way that the Wechsler Verbal scales focus on historical Gc. Nonetheless, such assessments do not illuminate a large aspect of intellectual functioning of adults well, in that they do not accurately describe the full range of intellectual capabilities of adults in occupational and avocational activities that are not common to the wider culture.

Of course, these various concerns still do not answer the question "what is adult intelligence?" Current measures of adult intelligence (the modern Wechsler Adult Intelligence Scale, now in its fourth edition; Wechsler, 2008) correlate reasonably well with measures of academic performance in college and beyond and correlate significantly, but not nearly as substantially, with measures of job performance. Other than these traditional criteria, it is not clear with what external criteria adult intelligence scores should be most highly associated. One could propose that an individual's ability to read and write; to memorize phone numbers; to solve real-world problems of finances, health, politics; and so on, should be the most appropriate indicators of adult intelligence. Demming and Pressey (1957), for example, suggested that in fact, the kinds of tests that compose Wechsler's Performance scale are not as relevant for adult intelligence as they are for the intelligence of children and adolescents. Demming and Pressey proposed that everyday activities (e.g., using a telephone directory, getting professional assistance) were much more suitable indicators of adult intelligence than traditional IQ test components. These researchers found that middle-aged adults performed better on such tests than younger adults and adolescents. Such investigations serve to point out that when assessing and understanding

adult intelligence, the underlying content of what makes up "adult intelligence" depends at least partly on the intended purpose of the assessment. Predicting occupational or educational success is a well-studied area. However, when it comes to deciding which adults are more intelligent than others outside of these contexts, the answer to what adult intelligence is, ultimately depends on one's theoretical viewpoints and applied purposes.

AGING AND LONGITUDINAL EXAMINATIONS OF ADULT INTELLIGENCE

In the years subsequent to the creation of the Wechsler scales and the development of theories of aging and adult intelligence, several important in-depth longitudinal studies of adult intelligence have been reported. A detailed description of these studies is beyond the scope of this chapter (for a review, see, e.g., Schaie, 1996). However, two main findings emerged from these studies. First, these investigations have pointed to the existence of substantial cohort differences in measured intelligence, with older cohort groups performing at lower levels on standardized tests of intellectual abilities. That is, on average, people born in earlier decades tend to perform less well on tests of intelligence than people born in more recent decades. The meaning of these findings is somewhat controversial (e.g., see Flynn, 1987, regarding the "secular rise" in intelligence scores), and various explanations have been proposed, such as differences in exposure to media (e.g., newspapers, television), nutritional differences, educational differences, and so on.

The second finding is that, although taking account of cohort differences results in attenuation of the steep age-related declines in intellectual abilities with increasing age in adulthood, peak levels on Performance/Gf-type tests are typically found for adults in their early 20s, with relatively steep declines after about age 30. Peak levels of Verbal/Gc tests (where the measured Gc is mainly historical Gc) occur a bit later than they do for Performance/Gf tests, but certainly by the late 30s and beyond, performance on these tests reaches a plateau and begins to decline at a slow rate as the individuals reach beyond their 40s. From an assessment perspective, the key inference from these findings is that any intelligence score for an adult that is not predicated on a comparison to his or her own age cohort is likely to be misleading from a normative perspective because the effects of aging will be confounded with mean age cohort differences.

CURRENT FRAMEWORKS FOR ADULT INTELLIGENCE

Extant theory and research on adult intelligence have developed substantially in the past few decades, in comparison to the early part of the 20th century. For the most part, the approaches can be divided into examination of Gf, examination of Gc, and integrated theory. Each of these foci is treated in turn here.

Gf

The examination of Gf-type abilities in adults and aging has developed into a set of investigations that range widely from descriptive studies of the age-related differences on several components and correlates of Gf to basic experimental investigations into the neurological correlates of Gf abilities. In the first type of investigation, researchers take a fundamentally top-down approach to understanding adult intelligence by focusing on the particular age-related patterns of Gf ability measures that make up adult intelligence tests. In both cross-sectional studies (e.g., Salthouse, 1994, 1996) and longitudinal studies (Baltes & Mayer, 1999; Schaie, 1996), investigators have determined that, by and large, speeded tests of abstract reasoning, short-term memory, math, and spatial abilities show varying degrees of decline with increasing age once the early 20s have been reached. Highly speeded tests, such as those of perceptual speed, and tests that tap complex spatial abilities appear to show the steepest age-related gradients, whereas tests of other Gf-related abilities show scores that are less steep but still declining with increasing age in adulthood, depending on the format and difficulty of the tests used to assess the particular abilities. As with the earlier studies, however, longitudinal investigations that follow the same individuals over a period of time show more

shallow gradients of change with increasing age, compared with cross-sectional studies that confound aging effects with cohort differences.

In the past 2 decades, the construct of working memory has been a popular source of investigation and discussion with respect to aging and adult intelligence. Working memory (e.g., see Baddeley, 1986) represents a specific set of short-term memory tasks that generally require the individual to keep multiple objects in active memory but also to manipulate or update the items in a complex fashion (e.g., counting the number of items-to-be-recalled at the same time that one is trying to remember the actual items). These tests appear to show even steeper declines with respect to increasing adult ages than other, more traditional, measures of Gf abilities. Theorists have proposed that a variety of mechanisms may be responsible for these effects, such as the amount of "neural noise" increasing with advanced ages in adulthood. Such efforts, as well as those that focus on identification of brain activity or the locations of cortical activity when examinees perform these tasks, follow an essentially bottom-up approach to understanding Gf. For the most part these efforts have not produced measures that are validated against any external criteria (e.g., occupational performance or competence in everyday intellectual activities). Even examination of elementary cognitive tasks has done little to illuminate the building blocks for individual differences in Gf abilities (e.g., see Carroll, 1980).

Several researchers of working memory and others have searched for pure measures of Gf, in the hope that they can better understand the nature of age-related changes in adult intelligence. Various measures have been proposed over the past 70 or more years to uniquely assess Gf. Prominent among these is Raven's Progressive Matrices task, a spatial inductive reasoning test. Researchers have decomposed the task into underlying strategies and sources of difficulty (e.g., Carpenter, Just, & Shell, 1990). Again, however, these investigations have not had much effect on the practical assessment of intellectual abilities in adults, although it should be noted that recent revisions of the Stanford–Binet scales now incorporate measures of working memory (e.g., see Roid, 2003).

Gc

Investigations of adult intelligence with respect to Gc primarily fall into the historical Gc domain. With the exception of those who claim that Gf or working memory cover the entire construct of adult intelligence, all modern assessments of adult intelligence include measures of Gc-type abilities (e.g., vocabulary, comprehension, fluency, general information). Different intelligence tests have a greater or lesser emphasis on verbal/Gc content. For example, the latest revisions of the Stanford–Binet scales tend to have a somewhat greater emphasis on verbal content, whereas the Wechsler scales tend to be more balanced between verbal and performance content. Intellectual ability tests for adults that are administered in group settings (e.g., the Wonderlic test; Wonderlic & Associates, 2002) have substantial Gc content, and similarly to the Army Alpha test, also require examinees to be able to read and write, which adds a somewhat greater demand for Gc-type abilities than the one-on-one Stanford–Binet and Wechsler intelligence tests.

There also have been a small number of investigations of adult intelligence that focus on current Gc, especially in terms of a wide variety of domain knowledge tests. Ackerman and his colleagues have examined age differences in several domains, such as science, humanities, business, law, health and nutrition, and current events (e.g., see Ackerman, 2000; Ackerman & Beier, 2006; Ackerman & Rolfhus, 1999; Beier & Ackerman, 2001, 2003). Although these investigations were cross-sectional studies (where potential cohort differences could be confounded with aging effects), with the exception of domain knowledge in the physical sciences (e.g., physics, chemistry), adults between 40 and 60 performed better, on average, than young adults between 18 and 25. If cohort effects in domain knowledge are similar to those for Gf and historical Gc, one would predict that middle-aged adults, on average, have much higher Gc than young adults, if we give them credit for what they know beyond common cultural content (i.e., historical Gc), at least up to about age 60 or 70. In addition, assessing adults on even a dozen or more Gc domains may barely scratch the surface of the vast domain knowledge and skills that adults have accumulated within

their intellectual repertoire. The approach that focuses on current Gc comes much closer to Binet's description of the "pedagogical method" of assessing intelligence than previous efforts for the assessment of adult intelligence, because this approach attempts to assess the depth and breadth of an individual's knowledge and skills, even if they are not common to the wider culture. In this way, one could differentiate between the capabilities of two different carpenters on the basis of the differences in their respective domain knowledge and skills in carpentry and also on the basis of their respective knowledge and skills in other domains, such as music, art, computer programming, meteorology, and so on. Whether it is possible to equate differences in *depth* of knowledge in a single area (e.g., for a specialist) with differences in *breadth* of knowledge across many areas (e.g., for a generalist, or a widely read specialist) remains an open question. Nonetheless, the external validation for such measures would be a catalogue of the kinds of intellectual activities that an adult could successfully complete in a variety of different domains, including those within and outside of the individual's occupation.

Integrated Theory

In the past few decades, theories of adult intelligence have coalesced around modifications to the Cattell approach, with some refinements introduced by Horn and Cattell (Horn, 1968, 1989; Horn & Cattell, 1967), and additional refinements by Carroll (1993). The consensus view of adult intelligence has become known as the C-H-C perspective to represent the amalgamation of approaches and empirical data from Cattell, Horn, and Carroll (e.g., see McGrew & Flanagan, 1998). From the 1950s through the 1980s, Cattell and Horn introduced additional ability factors to the Gf-Gc theory, such as general speediness (Gs), general visualization (Gv), and tertiary storage and retrieval (TSR), which involves long-term memory (see, e.g., Horn, 1989, and Horn & Noll, 1997). From their analyses and from existing cross-sectional and longitudinal data, they determined that, during adulthood, there are substantial declines in Gf, Gs, and Gv but substantial gains in TSR, along with Gc, at least up to about age 60. Carroll (1993), in his massive reanalysis of data from hundreds of abilities studies, concluded that general intellectual ability can be subdivided into eight broad factors of abilities; namely, Gf, Gc, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and processing speed (see Carroll, 1997). These ability factors can be further subdivided into abilities that represent narrower aspects of the underlying process and content associated with the higher order mental activities.

The C-H-C Framework and Assessment of Adult Intelligence

Two batteries of intelligence tests for adults that are aligned, to a greater or lesser degree, with the C-H-C framework are the Kaufman Adolescent and Adult Intelligence Test (KAIT; e.g., see Kaufman & Kaufman, 1997) and the Woodcock–Johnson Test of Cognitive Ability (e.g., see Woodcock, 1997). The KAIT can be used to obtain an examinee's overall IQ score and separate scores for Gf and Gc abilities. The Woodcock–Johnson also provides an overall IQ score, in addition to 20 cognitive ability "clusters" that correspond to Gf, Gc, and several other abilities outlined in the C-H-C framework (e.g., Gq [quantitative ability], Ga [auditory processing], and Gs [processing speed].) Both the KAIT and the Woodcock–Johnson overall IQ scores correlate substantially with scores obtained on the Wechsler scales, but they differ to some degree. Gc indexes and the Wechsler Verbal IQ tend to be most highly correlated, with measures of Gf and the Wechsler Performance IQ scores generally showing somewhat lower correlations. These differences are concordant with the different conceptualizations of the construct of adult intelligence. For example, Wechsler's clinical approach resulted in abilities of reasoning and speed grouped in the Performance scales, but tests such as the Woodcock–Johnson provide separate scores for Gf from Gs. On the one hand, these newer tests have not reached the widespread use of the Wechsler tests, and, to date, it is not clear whether the new tests provide incremental utility for the purpose of clinical diagnosis. On the other hand, the KAIT and Woodcock–Johnson tests provide a much more extensive sampling of adult intellectual

abilities than the Wechsler tests, so, for purposes of normal assessment, they may be expected to yield more extensive profiles of an individual's intellectual abilities.

Challenges for Assessing Adult Intelligence

Although modern intelligence assessment was first developed over 100 years ago, there remain significant challenges to the measurement and interpretation of adult intelligence. A few of the major challenges include issues of constancy of the IQ; conditions of testing, namely, maximal versus typical effort; and decisions on whether to assess broad aspects of adult intelligence versus measurement of narrow aspects of adult intelligence. Each of these issues is discussed briefly here.

Constancy of the IQ in adulthood. Although the IQ concept was not introduced by Binet (rather, it was proposed by Stern, 1914, and later implemented by Terman, 1916, in his translation and refinement of the Binet–Simon scales), there was an explicit assumption that intelligence as expressed by the IQ was a fixed trait. Once an individual received an IQ from an intelligence test, its value was expected to be constant except for fluctuations that were attributable to measurement error and specific qualities of particular assessment instruments. Indeed, this idea is reflected in many lay conceptualizations of IQ—once one has a score from childhood, it is carried forward through adulthood. Empirical research indicates that, if an IQ is obtained after a child reaches about age 6, its value remains reasonably consistent for most people, up to early adulthood (for a review, see Thorndike, 1940), although there are many instances of dramatic changes in IQ among some people in the population at large (e.g., see Bayley, 1949).

Once adulthood is reached, it is clear that IQ constancy in normative terms may be high (i.e., many people keep their relative standing on intelligence scores, with respect to their particular age cohort group), but absolute scores on individual intelligence tests may change markedly with increasing age, even when the assessment instruments are constant. Thus, it is especially important to consider stability and change in intelligence within the context of either relative standing or absolute scores.

Typical effort versus maximal performance. One of the key procedural requirements for intelligence assessment introduced by Binet, and carried forward to all later intelligence assessments, is that the examinee is encouraged and expected to put forth maximum mental effort to the task of answering questions and solving problems during the assessment. Binet reasoned that, to determine what the individual was capable of, the examinee must devote all available attention to the task. If an examinee was not interested in the assessment situation, or just poorly motivated to engage in intellectual tasks, the IQ computed from the test results could be expected to reflect an underestimation of the examinee's underlying intellectual ability. One potential problem inherent to this approach is that one often wants to assess intelligence to predict not just scores on other maximal performance indicators (e.g., the SAT or GRE scores) but rather intellectual accomplishments of a more typical nature, such as job performance over a period of a year, the number of novels written, or patentable discoveries made over a 10-year period. These indicators of intellectual accomplishment can be reasonably expected to depend not just on one's maximal performance but, more important, on the level of intellectual efforts made over extended periods of time—that is, one's typical intelligence.

There is no easy way out of this conundrum. By the time a person has reached adulthood in the modern industrialized society, the examiner need not remind the examinee to try hard on the test. Through long experiences with mental ability tests for grades and for educational or occupational selection, the examinee knows implicitly that a test means that he or she must put forth maximal effort. As such, one cannot expect useful assessment data, even if the examinee was instructed to just "give this test the kind of effort and attention that you usually devote to intellectual tasks." Tests of Gf, partly because they often depend on speed and abstract reasoning with novel tasks, are expected to be most

affected by fluctuations of effort and attention. Therefore, the theoretical difference between an individual's maximal performance and typical performance is likely to be largest for Gf-type abilities. Tests of Gc, however, are less, but not entirely, susceptible to differences between maximal and typical effort, for two reasons. First, the retrieval of information from long-term storage is much less effortful than, for example, solving novel word-problem math items. Second, and perhaps more important, when one considers current Gc (in contrast to historical Gc), one cannot acquire a substantial repertoire of domain knowledge if typical intellectual engagement with the environment is low. Individual adults who invest their intellect in acquiring knowledge over long periods of time are expected to be able to retrieve a much larger amount of domain knowledge from memory than those who do not make such investments. Maximal effort, in terms of highly focused attention during the assessment itself, is expected to yield relatively little gain in performance if the individual does not already have the knowledge or skills in his or her repertoire (see Ackerman, 1994).

Broad versus narrow assessments. The vast majority of global intelligence assessments are conducted for educational and occupational purposes related to children, adolescents, and young adults. Historically, assessment of global intelligence for middle-aged and older adults was conducted for clinical diagnostic purposes (e.g., to aid in the diagnosis of strokes, tumors, or dementia; see Chapter 9 of this volume for a more complete discussion of neuropsychological assessment) or forensic purposes (e.g., determining competency). Such tests were an important part of a psychometric battery (e.g., the Halstead–Reitan Battery; see Reitan & Wolfson, 1985) before the introduction of computerized tomography scans, and magnetic resonance imaging equipment, which now enable the neurologist to pinpoint the physical manifestations of neurological incidents. Although clinicians use adult intelligence assessments much less frequently for such diagnostic purposes, behaviorally based measures of intellectual abilities remain a critically important tool for prediction of the individual's ability to function in society. Tests such as the Mini-Mental State Exam (Folstein, Folstein, McHugh, & Fanjiang, 2001) are used to aid in the determination of an individual's overall intellectual competency. However, these measures are not primarily designed to rank order individuals in terms of their relative standing on intellectual abilities; rather, the tests are used to determine whether the individual has a threshold level of intellectual functioning that will allow him or her to remain independent, or require assistance.

Several narrow measures of adult intellectual functioning have been developed and refined over the past few decades, such as the Wechsler Memory Scales (Wechsler, 1997). These measures may focus on the diagnosis of specific impairments, such as with learning or short-term memory, working memory, fluency, and motor coordination. When used in conjunction with previous estimates of intellectual functioning, they can be used to aid in the diagnosis of particular kinds of mental decline (e.g., mild cognitive impairment, Alzheimer's disease, frontotemporal dementia) or the effects of some kind of head injury. Use of narrower measures is much better suited to differential diagnosis than omnibus intelligence tests and also can be used to make predictions about the time course of particular kinds of impairments.

FINAL OBSERVATIONS

Binet presented his approach to assessing intelligence in children as just one of three different methods of assessing intelligence—the psychological method. His preference for this method, especially in comparison to the pedagogical method of assessing intelligence, stemmed from his desire to remove, at least as much as possible, the influence of different socioeconomic backgrounds from the assessment of an individual's intelligence. His approach achieved limited success on this issue, in that literacy is not required for performing well on the Binet-type scales of intelligence. However, measures of socioeconomic status often demonstrate moderate correlations with individual differences in IQ.

Binet's method of assessing intelligence worked very well for predicting academic success in children (correlations [*r*s] between intelligence test scores and grades in school typically reach levels from .40 to .75; see Anastasi & Urbina, 1997). Upward revisions of the Binet scales and refinements to assessment of adult intelligence yield significant correlations with adult occupational performance (e.g., see Ackerman & Humphreys, 1991), although such correlations rarely exceed .50. For Binet, intelligence represented memory, imagery, imagination, attention, comprehension, suggestibility, aesthetic appreciation, moral sentiments, strength of will, and motor skill—all factors that are integral to school success.

Nearly 90 years ago, after the publication of results from the Army Alpha test, E. G. Boring suggested, not clearly tongue-in-cheek, that "intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test" (Boring, 1923, p. 35). Although such an operational definition works reasonably well to the degree that one can describe the component scales of an intelligence test (e.g., memory, vocabulary, reasoning), it does not tell us what kinds of scales we should include in an intelligence test if we were starting from scratch. Indeed, Wechsler (1939/1944) struggled with the decision of whether even to call his measure an "IQ" test, because IQ had been so highly associated with Binet-type scales—and there were clear differences in the content of Wechsler's tests, and even in the computation of IQ scores from the raw test scores, in comparison with the Binet-type scales.

Where Binet had the advantage of being able to use age differentiation and academic success as the key criteria for the evaluation of his intelligence test items, Wechsler and those who have followed him have found it difficult to obtain consensus on just what activities actually signify high or low intelligence in adults. Most investigators agree that adult intelligence measures should correlate positively with indicators of occupational performance, but there are also other nonability influences on such measures (e.g., motivation, years of experience, level of occupational complexity). One might also ask whether the traditional testing approaches to obtaining maximal effort on the part of the examinee are appropriate for predicting what intellectual accomplishments an individual is likely to achieve, because in many, if not most, adult activities, maximal engagement is perhaps less important than sustained typical engagement toward problem solving. Ultimately, whether one defines adult intelligence in terms of an individual's capability to quickly memorize and recall a new phone number, to complete one's tax forms without error, to perform well at work or in a family trivia game, to be able to complete crossword puzzles, to prepare and implement a menu for a large dinner party, or any of myriad other intellectual activities, will guide the assessment methods and interpretation of adult intelligence measures. By implication, such definition and assessment procedures will affect whether the resulting intelligence scores are meaningful for occupational success, the capability to perform other intellectual tasks, or determining success in everyday activities.

References

Ackerman, P. L. (1994). Intelligence, attention, and learning: Maximal and typical performance. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. 4. Theories of intelligence* (pp. 1–27). Norwood, NJ: Ablex.

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227–257. doi:10.1016/S0160-2896(96)90016-1

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *55*, 69–84. doi:10.1093/geronb/55.2.P69

Ackerman, P. L., & Beier, M. E. (2006). Determinants of domain knowledge and independent study learning in an adult sample. *Journal of Educational Psychology, 98*, 366–381. doi:10.1037/0022-0663.98.2.366

Ackerman, P. L., & Humphreys, L. G. (1991). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 223–282). Palo Alto, CA: Consulting Psychologists Press.

Ackerman, P. L., & Rolfhus, E. L. (1999). The locus of adult intelligence: Knowledge, abilities, and nonability traits. *Psychology and Aging, 14*, 314–330. doi:10.1037/0882-7974.14.2.314

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York, NY: Prentice Hall.

Baddeley, A. (1986). *Working memory*. New York, NY: Oxford University Press.

Baltes, P. B., & Mayer, K. U. (Eds.). (1999). *The Berlin Aging Study: Aging from 70 to 100*. New York, NY: Cambridge University Press.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology, 75*, 165–196.

Beier, M. E., & Ackerman, P. L. (2001). Current events knowledge in adults: An investigation of age, intelligence and non-ability determinants. *Psychology and Aging, 16*, 615–628. doi:10.1037/0882-7974.16.4.615

Beier, M. E., & Ackerman, P. L. (2003). Determinants of health knowledge: An investigation of age, gender, abilities, personality, and interests. *Journal of Personality and Social Psychology*, *84*, 439–447. doi:10.1037/0022-3514.84.2.439

Binet, A., & Simon, T. (1973). *The development of intelligence in children* (E. Kite, Trans.). New York, NY: Arno Press. (Original work published 1905)

Boring, E. G. (1923, June 6). Intelligence as the tests measure it. *New Republic*, pp. 35–37.

Brigham, C. C. (1922). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431. doi:10.1037/0033-295X.97.3.404

Carroll, J. B. (1980). *Individual difference relations in psychometric and experimental cognitive tasks*. (Tech. Rep. No. 163). Chapel Hill: University of North Carolina, The L. L. Thurstone Psychometric Laboratory. (NTIS No. AD-A 086057 and ERIC Document Reproduction Service No. ED 191 891)

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 29–120). Cambridge, MA: Cambridge University Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511571312

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122–130). New York, NY: Guilford Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, *40*, 153–193. doi:10.1037/h0059973

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York, NY: World Book.

Conrad, H. S. (1930). General-information, intelligence, and the decline of intelligence. *Journal of Applied Psychology, 14*, 592–599. doi:10.1037/h0069963

Demming, J. A., & Pressey, S. L. (1957). Tests "indigenous" to the adult and older years. *Journal of Counseling Psychology, 4*, 144–148. doi:10.1037/h0043284

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191. doi:10.1037/0033-2909.101.2.171

Folstein, M. F., Folstein, S. E., McHugh, P. R., & Fanjiang, G. (2001). *Mini-mental state examination*. Odessa, FL: Psychological Assessment Resources.

Hebb, D. O. (1939). Intelligence in man after large removals of cerebral tissue: Report of four left frontal lobe cases. *Journal of General Psychology, 21*, 73–87. doi:10.1080/00221309.1939.9710587

Hebb, D. O. (1942). The effect of early and late brain injury upon test scores, and the nature of normal adult intelligence. *Proceedings of the American Philosophical Society*, *85*, 275–292.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75*, 242–259. doi:10.1037/h0025662

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 61–116). New York, NY: Freeman.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107–129. doi:10.1016/0001-6918(67)90011-X

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: *Gf-Gc* theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York, NY: Guilford Press.

Hsiao, H.-H. (1927). *The performance of the Army Alpha as a function of age*. Unpublished master's thesis, Columbia University, New York, NY.

Jones, H. E., & Conrad, H. S. (1998). The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty. In M. P. Lawton & T. A. Salthouse (Eds.), Essential papers on the psychology of aging (pp. 149–174). New York: New York University Press. (Original work published 1933)

Kanfer, R., Ackerman, P. L., Murtha, T., & Goff, M. (1995). Personality and intelligence in industrial and organizational psychology. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 577–602). New York, NY: Plenum Press.

Kaufman, A. S., & Kaufman, N. L. (1997). The Kaufman Adolescent and Adult Intelligence Test.

In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 209–229). New York, NY: Guilford Press.

Lippmann, W. (1922, October 25). The mental age of Americans. *New Republic*, pp. 213–215.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston, MA: Allyn & Bacon.

Owens, W. A., Jr. (1953). Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs, 48*, 3–54.

Pintner, R., & Paterson, D. G. (1917). *A scale of performance tests*. New York, NY: Appleton. doi:10.1037/11199-000

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Roid, G. H. (2003). *Stanford–Binet intelligence scales* (5th ed.). Itasca, IL: Riverside.

Salthouse, T. A. (1994). The nature of the influence of speed on adult age differences in cognition. *Developmental Psychology, 30*, 240–259. doi:10.1037/0012-1649.30.2.240

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*, 403–428. doi:10.1037/0033-295X.103.3.403

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle Longitudinal Study*. New York, NY: Cambridge University Press.

Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin*, *70*, 671–680. doi:10.1037/h0026811

Stern, W. (1914). *The psychological methods of testing intelligence* (G. M. Whipple, Trans.). Baltimore, MD: Warwick & York.

Terman, L. M. (1916). *The measurement of intelligence*. Boston, MA: Houghton Mifflin. doi:10.1037/10014-000

Thorndike, R. L. (1940). "Constancy" of the IQ. *Psychological Bulletin*, *37*, 167–186. doi:10.1037/h0061268

Wechsler, D. (1944). *The measurement and appraisal of adult intelligence* (3rd ed.). Baltimore, MD: Williams & Wilkins. (Originally published 1939)

Wechsler, D. (1997). *The Wechsler Memory Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth edition. Technical manual*. San Antonio, TX: Pearson.

Wonderlic & Associates. (2002). *Wonderlic Personnel Test manual*. Libertyville, IL: Author.

Woodcock, R. W. (1997). The Woodcock–Johnson Tests of Cognitive Ability—Revised. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 230–246). New York, NY: Guilford Press.

Yoakum, C. S., & Yerkes, R. M. (Eds.). (1920). *Mental tests in the American Army*. London: Sidgwick & Jackson. doi:10.1037/11054-000

CHAPTER 9

ASSESSMENT OF NEUROPSYCHOLOGICAL FUNCTIONING

Antonio E. Puente and Antonio N. Puente

Ebbinghaus's statement, "Psychology has a long past but a short history," applies to clinical neuropsychological assessment. The earliest recorded work in neuropsychological testing goes back to the work of Franz well over a century ago. Indeed, a review of psychology's two foundational books, Wundt's *Textbook of Physiological Psychology* (1904, English translation) as well as James's *Psychology* (1890), makes it clear that neuropsychology and neuropsychological assessment have been central to the mission of psychology since its founding as a discipline. In an attempt to bridge philosophical questions to scientific methodology, Wundt applied the scientific process, and the outgrowth was neuropsychology. In a review of the total number of chapters of both Wundt's and James's books, the majority of the chapters discuss the use of "formal and informal" tests to understand the relationship of "psychic processes" to brain function. For example, Chapter 2 of James's book provides a delineation of processes titled "Functions of the Brain," and on page 20, a portion of that chapter is titled "General Notion of the Hemispheres."

Over the next 50 years, psychology drifted and operated from behaviorism as the main theoretical perspective to understand psychological processes. With the work of Watson, as outlined in *Psychology from the Standpoint of a Behaviorist* (1919) and, subsequently, Skinner, with *Science and Human Behavior* (1953), psychology became focused on understanding behavior primarily through the lens of behaviorism. The rise of clinical neuropsychology as a primary method for understanding human behavior and dysfunction over the past 30 years is a return to the roots of our discipline.

This chapter provides an overview of that return and a presentation of the basics of neuropsychological assessment and evaluation (terms that are used interchangeably in this chapter). After a brief historical overview, three sections are presented covering (a) clinical neuropsychology as a profession, (b) neuropsychological assessment, and (c) the future of neuropsychological assessment.

BRIEF HISTORY OF NEUROPSYCHOLOGICAL ASSESSMENT

Neuropsychological testing has a history of approximately half a century, although the first half was fraught with limited information, and the second has been marked by very rapid growth and is well chronicled. For example, the first article on the history of clinical neuropsychology was published by Goldstein in 1985. Since that time, approximately 20 articles have been published on the topic. Puente (1989, 2005), Reitan (1989), Fitzhugh-Bell (1997), Puente and Marcotte (2000), Zillmer (2004), and Hartlage and Long (2009) have provided some of the most comprehensive information about the history of clinical neuropsychology. A few authors (e.g., Reitan, 1989) have focused almost exclusively on specific testing, whereas others (e.g., Benton, 1972) have provided more generic overviews. Other important "historical" works, such as the acclaimed work by Kurt Goldstein from 1942, *Aftereffects of Brain Injuries in War*, described injuries and

DOI: 10.1037/14048-009
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology, K. F. Geisinger (Editor-in-Chief)

outcomes rather than processes used to understand them or a truly historical presentation.

Probably the first book to address neuropsychological assessment was Franz's *Handbook of Mental Examination Methods*, published in 1920 and based on practices he began around 1910. This book contains a series of lectures involving "neurological and mental examination methods" he presented to interns at the Government Hospital for the Insane in Washington, DC. Several mental tests were listed, and methods to address both time and observational information were found. Suggested to be effective both for diagnostic and research purposes, his methods included assessment for the following: sensation, movement, language, attention apprehension and perception, memory, association, calculation, and general intelligence.

Russell, Neuringer, and Goldstein (1970) published what could be considered the first book exclusively devoted to neuropsychological testing in English, *Assessment of Brain Damage: A Neuropsychological Key Approach.* The book was an outgrowth of work since Goldstein's dissertation in 1963 on testing for brain damage. Although a good portion of the foundations for this approach was directed to psychiatric populations, this book was significant in that it addressed the application of such tests to neurological patients—a focus that has been maintained within neuropsychology to the present. Additionally, it presented a systematic approach to determine brain dysfunction. All three authors were heavily influenced by the work of Reitan, a student of Halstead at Chicago. Reitan took tests such as the Seashore Rhythm Test from vocational and related fields and applied them to understand brain dysfunction. It was not until 1974, however, that Reitan himself, with Leslie A. Davison, finally published another landmark book on neuropsychological testing, *Clinical Neuropsychology: Current Status and Applications*. In collaboration with Davison, Reitan published an overview of his battery and clinical neuropsychology for the psychometrically based North American audience with some "norms."

Before the publication of his first book, there were only two methods for learning Reitan's approach (i.e., the Halstead–Reitan Neuropsychological Battery): study directly with him like the Reed brothers, or obtain the information from Reitan's workshops. The majority of individuals learned this method through the latter means. Typically, these colloquia were lengthy presentations of Reitan's ideas including theory, protocol, and application of a battery of tests. The only data available (e.g., normative information) on these tests were, for many years, presented at these workshops, and until the National Academy of Neuropsychology (NAN) annual conference in Orlando in 1988, only a small portion of clinical neuropsychologists had attended. Thus, although some understood Reitan's approach and battery, most practitioners were unable to appreciate the evolution of Reitan's thinking.

Although assessment of brain damage was increasing, there was relatively little written that was comprehensive in terms of using psychological tests rather than batteries. The works of Reitan as well as of Goldstein focused on a very limited approach. However, in 1972 while at the University of Iowa, Benton wrote a seminal chapter titled "Psychological Tests for Brain Damage," which presented a more comprehensive approach to understanding brain dysfunction using psychological tests. Benton suggested that an evaluation could include a variety of psychological tests rather than just a battery. From this perspective, a more robust and comprehensive understanding of the brain and the potential set of impairments could be achieved. Benton outlined the first reported survey of neuropsychological tests for adults and children, including measures of the following domains: general intelligence, reasoning, memory and orientation, language functions, perceptual and perceptuomotor performance, response speed and flexibility, and attention and concentration.

After this introduction of multiple tests came an era focusing on the application of those tests to understand specific syndromes. An excellent and early example of this approach appeared in Parsons and Butters's (1987) *Neuropsychology of Alcoholism: Implications for Diagnosis and Treatment*. This book, as an example of many others to this day (e.g., Goldstein, Incagnoli, & Puente, 2011), used the different approaches proposed by Reitan, Benton, and others to begin systematic analysis of specific syndromes.

The value of such descriptions has been based on the value of the neuropsychological instruments used to understand those syndromes. As a result, over the past 2 decades, an ever-expanding list of neuropsychological tests has appeared in the literature focusing on specific disorders.

CLINICAL NEUROPSYCHOLOGY

Clinical neuropsychology was formed as a result of scientific evolution and amalgamation of several disciplines (e.g., neurology and clinical psychology; Sperry, 1995). In 1996, after much work on the part of individuals such as Meier, clinical neuropsychology was formally recognized by the American Psychological Association (APA) as a specialty in psychology, joining the existing specialties of clinical, counseling, and school psychology (Boake, 2008). Clinical neuropsychology is a specialty that uses assessment and intervention to understand brain–behavior relationships and applies this knowledge to human problems (APA Commission for the Recognition of Specialties and Proficiencies in Professional Psychology, 1996). The fundamental goal of clinical neuropsychology is to determine psychological problems (e.g., behavior, cognition, and mood) affected by central nervous system dysfunction (Meier, 1997).

A clinical neuropsychologist is a professional within the field of psychology with expertise in the applied science of brain–behavior relationships (Barth et al., 2003). Neuropsychologists use expertise in brain–behavior relationships to assess, diagnose, and provide effective interventions (e.g., therapy and rehabilitation) for individuals of all ages with neurological, medical, and psychiatric conditions (APA Division 40 Executive Committee, 2006; Barth et al., 2003). Barth et al. (2003) stated, "The clinical neuropsychologist uses psychological, neurological, physiological, cognitive and behavior principles, techniques and tests to evaluate patients' neurocognitive, behavioral, and emotional strengths and weaknesses and their relationship to normal and abnormal central nervous system functioning" (p. 554). Clinical neuropsychologists are practitioners; have a doctoral degree from an accredited university program; completed an internship in professional psychology, which is equivalent to 2 years of full-time specialized training at the postdoctoral level in the field and practice of clinical neuropsychology; and have a license to practice psychology in their respective state or province or are employed as neuropsychologists by an exempt agency (Barth et al., 2003).

Neuropsychologists engage in several professional activities, but neuropsychological assessment accounts for the largest amount of professional time (Rabin, Barr, & Burton, 2005; Sweet, Peck, Abramowitz, & Etzweiler, 2002). The ontogeny of clinical neuropsychology is suggested to be due to its utility in localization, lateralization, and lesion detection—the so-called "three Ls" (Hartman, 1991). This contribution was accomplished with comprehensive assessments, which included mood, cognitive, personality, and behavioral instruments.

The advent and improvement of neuroimaging have decreased the necessity of neuropsychological evaluations for the three Ls (Beaumont, 2008; Marcotte, Scott, Kamat, & Heaton, 2010). Nonetheless, these technological advancements have not made clinical neuropsychology obsolete; rather, they have refined its purpose. Lezak, Howieson, and Loring (2004) have suggested that neuropsychological assessments are often obtained for the following:

- diagnosis,
- patient care,
- treatment planning,
- treatment evaluation,
- research, and
- forensics.

Historically, neuropsychological assessments were the most frequently sought for assistance with diagnostic concerns and remain the most frequent referral question (Marcotte et al., 2010). However, the improvement of neurodiagnostic techniques has decreased the need of neuropsychological assessment for diagnosis (Beaumont, 2008; Lezak et al., 2004). Nonetheless, the use of neuropsychological assessment as a diagnostic method is frequently used in differential diagnosis, often to distinguish between psychiatric and neurogenic and between different neurological conditions (Lezak et al., 2004; Meier, 1997) as well as to determine possible localization of dysfunction (Tonkonogy & Puente, 2009).

Neuropsychological assessment allows for an in-depth analysis of functional limitations associated with brain dysfunction and is required for diagnosis by some diagnostic criteria for neurological disorders such as Alzheimer's disease given that biomarkers are not yet reliable (McKhann et al., 1984; Storey, Slavin, & Kinsella, 2002).

Utility of neuropsychological assessment is not limited to clinicians but also benefits academicians and others interested in research. The use of neuropsychological assessments for this purpose is often attributed to Halstead because he is credited with applying the "test battery" approach to investigate brain–behavior relationships of normal and brain-damaged participants in a systematic and standardized format (Reitan, 1994). Neuropsychological assessments are frequently used in research to better understand the effects of mood disorders (Porter, Bourke, & Gallagher, 2007), psychotic disorders (Palmer, Dawes, & Heaton, 2009), neurodegenerative diseases (Libon et al., 2007), physical conditions (e.g., hypertension; Elias, Elias, Sullivan, Wolf, & D'Agostino, 2003), and psychological and medical treatments (e.g., psychotherapy, chemotherapy, and heart surgery; McClintock, Husain, Greer, & Cullum, 2010; Tully, Baker, Knight, Turnbull, & Winefield, 2009; Vardy, Rourke, & Tannock, 2007). Additionally, using neuropsychological assessment for "basic" research also helps develop new assessment techniques and instruments as well as norms that help to increase sensitivity and specificity of neuropsychological dysfunction (Ostrosky-Solís, Ardila, & Rosselli, 1999).

Assessing neuropsychological functioning in clinical settings has proven increasingly beneficial and common, and the use of neuropsychological assessment in forensic settings has become increasingly valuable (Horton, 2010). In contrast to clinical neuropsychology, assessment in forensic settings often has different goals, questions, clients, and techniques (e.g., the decision-making process; Prichard, 1997; see also Chapter 16, this volume). Regardless, neuropsychological evidence in forensic settings assists third parties (e.g., judges and juries) in making just legal decisions (Horton, 2010) and has been elaborated extensively by Sbordone (e.g., Sbordone & Saul, 2000) as well as by McCaffrey and colleagues (McCaffrey, Williams, Fisher, & Laing, 1997).

NEUROPSYCHOLOGICAL ASSESSMENT

Although there is variability in how neuropsychological assessments are conducted, the basic purpose is to acquire, analyze, and integrate neurological and neuropsychological data from multiple sources (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Typically, a neuropsychological assessment involves records review, interview, testing, and report writing. Records possibly provide the neuropsychologist with a general idea of what the presenting problem will be, and the interview is vital to gather a large and varied amount of data and clarify uncertainties in the clinical record and initial presentation. Testing involves the administration of various procedures and measures to patients and is based on and follows record review and interview.

The following sections provide an overview of the basic elements and processes comprising neuropsychological assessments. Although there is one assessment approach that is more prevalent and favored among neuropsychologists, both major approaches (i.e., fixed and flexible battery) are discussed in the sections that follow as well as the measures and norms associated with these procedures. After this discussion, some general considerations that affect neuropsychological assessment are examined. For example, technicians have become particularly important in neuropsychological assessments; therefore, this topic deserves attention. Finally, norms and time spent in a neuropsychological evaluation is covered, and the neuropsychological report is briefly explained.

Records Review

With records acquired before and information gathered during the interview, the neuropsychologist is able to develop hypotheses and administer tests to confirm or disprove various working hypotheses (Yochim, 2010). Typically, records are the first type of information available. However, in some cases, especially Social Security disability determination cases, in which few or no records are available. In these situations, important records (e.g., educational) are not available, as is often the case when

the individual has attended school in other countries. The reasons are varied, but typically such records are difficult to obtain, the number of personnel available to obtain such records is low, and the impact that the case carries is "limited" (as opposed to, e.g., death penalty cases).

In contrast, numerous challenges remain when records are obtained or available. In the example of an individual who has been educated in another country, the records are hard to equate to the United States, as educational attainment is not equivalent across countries. In other cases, such as complex forensic ones, large amounts of information are often available through extensive historical and laborious mining of records, sometimes done by mitigating "experts" such as paralegals, case managers, and social workers as well as neuropsychologists. Given these difficulties, historical information based on records is not always included in neuropsychological assessments even though it may comprise important and useful information.

Historical information outlined by existing records provide a wealth of data about past and present status, but synthesis of that information is necessary and can be challenging (Howieson & Lezak, 2010). This synthesis is best incorporated in a narrative format as a part of the written report of the neuropsychological assessment or in tabular form. A table can visually summarize the salient points critical to that synthesis and provide a trajectory of neurobehavioral changes over time. The synthesized records provide the professional with a contextual framework of the client and allows for preliminary hypotheses about client's difficulties to be pursued in subsequent testing. However, having the records before the evaluation may bias the evaluation procedure as well as the findings and interpretation of evaluation data. Although records review could increase the likelihood of bias into the neuropsychological evaluation, this qualitative information often provides the professional with the most representative context for the individual's presenting problems. In addition to the interview, it is almost always used to determine the necessary neuropsychological procedures to implement that challenge the clinician's hypotheses. Generally speaking, the goal of the record review is to place the individual within a socio-historical-cultural context as a means of providing baseline information about neuropsychological functioning (Luria, 1973, 1980).

Interview

The interview, or neurobehavioral status exam, in a neuropsychological evaluation is critical, given that it provides information for two of the components of an evaluation (i.e., history and behavioral observations). Although the neuropsychologist may have an accurate understanding of the client's functioning from records received previously, conducting an interview before testing is imperative to determine whether testing is necessary and, if so, what types of tests should be administered (Yochim, 2010). For example, if the client is heavily medicated, actively psychotic, or physically unable, then testing may be inappropriate—or even unethical—because significant error would be introduced (Vanderploeg, 2000). The interview also provides the clinician with an initial understanding of the level of cooperation of the client and what, if any, accommodations or modifications for the evaluation are warranted (Strauss, Sherman, & Spreen, 2006).

Interviews differ among practitioners, typically in the amount of structure implemented and interpersonal style (for further discussion of structured and unstructured interviews, please refer to Chapter 7, this volume). Although there is variability, there are standard areas to cover in a neuropsychological interview, including demographics, medical/health, developmental, educational, social, and occupational history as well as current medical/health status and the effect of the disorder on the client's life (Strauss et al., 2006; Yochim, 2010). The interview provides the neuropsychologist a chance to educate the client about the evaluation and addresses client concerns. In essence, the interview is both a data-gathering activity and an educational one.

Interviewing is not limited to the client but also includes significant others, children, parents, and even teachers and employers. If the availability arises, structured affidavits in forensic cases may be of value, especially in understanding premorbid levels of functioning and descriptions of more ecologically valid behaviors. Collateral interviews are best conducted without the patient being present to

enhance the validity of the information provided. It may be worthwhile to ask similar questions of the collateral interviewee and the patient to glean the patient's understanding of his or her difficulties.

Although interviews are typically semistructured, structured interviews are sometimes implemented to ensure that certain required information is obtained (Rogers, Bagby, & Dickens, 1992). This approach increases the likelihood that important information is included and that replication of the interview is more easily achieved. In contrast, unstructured interviews permit a glimpse into the patient's ability to develop themes and organize his or her thoughts, and they allow for additional information to be gathered about the individual's condition. Structured interviews are probably most beneficial for clinicians with less experience and in forensic cases where the obtained information will become available to a third party. At the same time, increased time and lack of fluidity may hamper the gathering of sensitive or subtle information. Ultimately, the most important aspect of the interview is to allow the clinician to formulate working hypotheses about a client's condition and implement measures to test his or her ideas as the interview sets the foundation (e.g., "medical necessity") for testing.

Testing Approaches

Generally speaking, there are two major approaches to neuropsychological assessment: fixed battery and flexible battery. The fixed battery, or standardized battery, approach uses the same battery of tests for every client, despite different presenting difficulties and referral questions (Fennell, 2000). The flexible battery approach uses a core battery of tests and techniques for clients with various syndromes (e.g., dementia and traumatic brain injury; Sweet et al., 2002). In contrast to fixed battery approaches, the tests implemented vary based on practitioner as well as context (e.g., inpatient/outpatient setting and syndrome). Given this variability, how tests are organized as well as the most frequently used tests in different contexts are discussed next.

Fixed battery. Two well-known fixed batteries are the Halstead–Reitan Neuropsychological Test Battery (HRNTB) and the Luria-Nebraska Neuropsychological Battery (Golden, Hammeke & Purisch, 1978; Reitan & Wolfson, 2004). The HRNTB is the most researched and used fixed neuropsychological test battery (Horton, 2008; Reitan & Wolfson, 2004). The HRNTB is based on the ideas of Halstead, who believed that there were two types of intelligence: psychometric and biological (Reynolds, Castillo, & Horton, 2008). Psychometric intelligence is what is measured by intelligence tests (e.g., Stanford–Binet), whereas biological intelligence reflects the adaptive abilities of a healthy central nervous system (Reitan, 1994; Reynolds et al., 2008). To determine biological intelligence, Halstead selected 13 tests, given that the brain-damaged individuals whom he examined had a wide range of deficits, and traditional intelligence tests were not always sensitive indicators of brain damage; some patients with significant damage did not exhibit deficits in functioning (Reitan & Wolfson, 2004; Reynolds et al., 2008). The HRNTB's frequency of use is attributed to the empirical evidence of its ability to evaluate brain-damaged individuals accurately as a battery, given both its comprehensive nature and its superior sensitivity for subtle deficits (Horton, 2008). It has evolved to distinguish accurately between normal and brain-damaged individuals, and, given that patients had a wide range of deficits, it was necessary to include numerous tests to examine these difficulties adequately (Reitan & Wolfson, 2004).

The battery of tests that constitute the HRNTB has been modified, as Reitan has added and removed several tests to improve the sensitivity to damage of the central nervous system (Reitan & Wolfson, 2004). Currently, the HRNTB includes 10 tests: the Speech–Sounds Perception Test (SSPT), Rhythm Test, Reitan–Indiana Aphasia Screening Test (AST), Tactual Performance Test (TPT), Tactile Form Recognition Test, Sensory-Perceptual Examination, Grip Strength Test, Finger Tapping Test, Category Test, and Trail-Making Test (TMT; Reitan & Wolfson, 2004). When the HRNTB is administered, the neuropsychologist may also include a traditional measure of intelligence (e.g., Wechsler Adult Intelligence Scale [WAIS], 4th edition) as well as a measure of academic achievement (e.g., Wide Range Achievement Test) and an objective personality

inventory such as the Minnesota Multiphasic Personality Inventory (MMPI; Horton, 2008).

The SSPT consists of 60 nonsense words with an "ee" sound presented on a recording and requires the individual to indicate which sound they heard out of four choices on an SSPT answer sheet (Reitan & Wolfson, 2004). It measures auditory memory, rhythmic discrimination, and attention ability; is designed to be relatively easy; and is one of two measures that evaluate the first level of central processing. The second measure in the HRNTB that measures the subject's attentiveness (i.e., first level of central processing) is the Rhythm Test (Reitan & Wolfson, 2004). Thirty pairs of rhythmic beats are presented to the client from a recording, and the individual is requested to determine whether the beats are the same or different. Although this test measures the client's attention, it specifically evaluates auditory perception and nonverbal auditory discrimination.

The AST measures different language functions, including naming, spelling, reading, writing, enunciating, identifying numbers and letters, and simple arithmetic (Reitan & Wolfson, 2004). This test identifies expressive or receptive language deficits, which is determined by the amount and type of errors committed (Johnson & D'Amato, 2011). In contrast to the AST, the TPT is a nonverbal test that examines an individual's ability to place 10 geometric blocks into 10 matching spaces on a board slanted 45° while blindfolded (Horton, 2008). The test is first performed with the subject's dominant hand, followed by the nondominant hand and, finally, both hands (Reitan & Wolfson, 2004). The time needed to complete each trial and errors are recorded and are interpreted to determine one's complex problem-solving skills. After completion of the task with both hands, the blindfold is removed and the examinee is requested to draw as many shapes as they can remember and place them in the accurate location. The number of correct shapes remembered and accurate location provide separate scores that can be used as measures of spatial learning (Horton, 2008).

Albeit similar, the Tactile Form Recognition Test is a separate test in the HRNTB that measures a client's ability to distinguish shapes by touching with their hands (Reitan & Wolfson, 2004). A board blocks the client's hand, and the client is requested to identify flat plastic shapes. The test is completed for both hands, and although other functions are involved, it is suggested to provide information about the contralateral parietal area and is a sensitive measure of brain damage (Reitan & Wolfson, 2004).

The Sensory-Perceptual Examination, (i.e., Reitan–Klove Sensory-Perceptual Examination) is a standardized and adjusted version of a behavioral neurologist's examination measuring the visual, auditory, and tactile sensory functions of the central nervous system (Horton, 2008; Reitan & Wolfson, 2004). Another basic ability, motor strength, is evaluated during the Grip Strength subtest. Grip strength is assessed with a hand dynamometer, the individual is requested to use each hand twice, and the mean score is recorded. Finger Tapping, a measure of motor speed, requires the client to press a lever attached to a small board and a counter as quickly as possible for 10 seconds with each hand on five consecutive trials.

In contrast to motor and sensory abilities, abstraction and problem solving are measured by how quickly the client is able to complete the TPT as well as the Category Test and the TMT (Reitan & Wolfson, 2004). The Category Test comprises seven subtests with a total of 208 items, requiring a client to choose the correct response out of four possibilities based on the principle of that particular set (Strauss et al., 2006). The client must deduce the underlying principle from the subtest with the feedback they received from their choices, as the examiner is not permitted to provide cues; rather, the examiner informs the client if the response is correct or incorrect. Originally, the Category Test was presented by means of a slide projector, but booklet and computer adaptations are now available (Strauss et al., 2006).

Although the Category Test is still widely used as a measure of abstract reasoning and problem solving, the TMT is more frequently administered (Ojeda & Puente, 2010). The TMT consists of two parts, A and B (Reitan & Wolfson, 2004). Trails A requests the client to draw lines that connect circles in numerical order from 1 to 25, whereas Trails B

requires the client to connect 25 circles by alternating between numbers and letters in sequence. The client is instructed to complete this task as quickly as possible. Errors are indicated by the examiner, and the examinee is redirected to the previous position (Reitan & Wolfson, 2004). The time taken to complete and errors produced generate separate scores and provide sensitive measures of cerebral functioning, and more specifically, frontal lobe functioning (Demakis, 2004).

The HRNTB provided an avenue and example for other neuropsychological test batteries to follow, such as the well-known and frequently administered Luria-Nebraska Neuropsychological Battery (Golden, 1982). The Luria-Nebraska Neuropsychological Battery, previously known as the Luria-South Dakota Neuropsychological Battery, evolved from the methods of Russian neuropsychologist, Alexander Luria (Goldstein, 2000). He endorsed qualitative procedures and was regarded as an intuitive genius, and he operated from deduction to determine the underlying deficit of an individual syndrome using a functional system approach (Golden, 1982). Luria was a renowned clinician and theorist, but his neuropsychological procedures were not standardized. Although controversial, Golden et al. (1978) standardized and validated Luria's procedures, which provided practitioners a comprehensive test battery built on his procedures. This battery now is supported by numerous empirical investigations and is widely administered by neuropsychologists (Goldstein, 2000). Although not as frequently used as previously, the battery allows for the development of a deficit analysis and an alternative fixed battery.

Golden and colleagues developed two forms of the Luria-Nebraska Neuropsychological Battery: Form I in 1980 and Form II in 1985 (Golden et al., 1978; Golden, Purisch, & Hammeke, 1985). Both have the same theoretical basis as they are a combination of Luria's qualitative procedures, with standardized and quantitative methods. These forms have separate administration materials but share 84 items in common. Form I has 269 items and 11 clinical scales, whereas Form II has 279 items and 12 clinical scales (Walker et al., 2008). The current battery takes approximately 1 1/2 to 2 1/2 hours to administer, which is considered an improvement, as it shorter than HRNTB (Golden et al., 1985).

Items are scored on a 3-point scale; 0, 1, and 2 indicate normal, borderline, and abnormal performance, respectively. Individual items are summed for each clinical scale and converted to T scores with a mean of 50 and standard deviation of 10 (Golden et al., 1978; Goldstein, 2000). The 12 clinical scales that make up Form II include the original 11 clinical scales plus Immediate Memory (Goldstein, 1985). The 11 original clinical scales are: Motor Functions, Rhythm, Tactile Functions, Visual Functions, Receptive Speech, Expressive Speech, Writing, Reading, Arithmetic, Memory, Intellectual Processes, and Immediate Memory.

The Motor Functions scale measures the ability to plan and complete simple motor abilities of the upper extremities and the face. This scale is similar to a standard neurological exam. It is the longest of the 12 clinical scales and organized for one to understand motor activity as a complex functional system (Golden, 1982). The Rhythm scale also requires motor abilities; however, it measures the ability to perceive and comprehend tones and rhythmic patterns accurately by requiring the client to reproduce words or rhythms or discriminate between tones. The Tactile Functions scale examines cutaneous and proprioceptive functions such as localizing touch, discriminating between two points and various degrees of pressure, perceiving the direction of a moving stimulus, and identification of various figures. Another sensory function thoroughly examined in the Luria-Nebraska Neuropsychological Battery is vision, evaluated with the Visual Functions scale. Golden (1982) indicated that this scale is "designed to evaluate a wide range of visual functions and is thus highly sensitive to right hemisphere dysfunction as well as dysfunction in posterior portions of the brain" (p. 60).

Comprehending and producing speech is measured by the Receptive Speech and Expressive Speech clinical scales. The examinee is required to choose pictures or verbal descriptions of what they heard on the Receptive Speech scale, whereas fluency and articulation ability is examined on the Expressive Speech scale by requiring the client to read and repeat verbal information (Walker et al.,

2008). The Writing scale evaluates an examinee's spelling, copying, and writing on a basic level. Similarly, the Reading scale examines basic reading ability by requesting the client identify sounds and read letters, words, sentences, and paragraphs. Fundamental and simple arithmetic skills such as calculation are examined on the Arithmetic scale, and the ability to encode and learn verbal and nonverbal information is measured by the Memory scale. The Intellectual Processes scale evaluates reasoning within different frameworks and contains similar items to measures of intelligence (Golden et al., 1985). The last clinical scale, Intermediate Memory, examines retrieval and maintenance of previously presented information.

Information can be organized into summary, localization, and factor scales using data obtained from the 12 clinical scales (Golden et al., 1985). There are five summary scales: Pathognomonic, Right Hemisphere, Left Hemisphere, Profile Elevation, and Impairment. The Pathognomonic scale contains items infrequently missed by healthy individuals and is sensitive to brain dysfunction (Goldstein, 2000). The Right Hemisphere and Left Hemisphere scales comprise items evaluating tactile and motor functioning of the respective side of the body. Profile Elevation and Impairment evaluate present functioning and degree of dysfunction, respectively (Tsushima, 2010).

As there are five summary scales, there are eight localization scales to best infer location of brain damage. The localization scales include Left Frontal, Left Sensorimotor, Left Parietal-Occipital, Left Temporal, Right Frontal, Right Sensorimotor, Right Parietal-Occipital, and Right Temporal (Golden et al., 1985). The factor scales comprise items representing different neuropsychological functions (Walker et al., 2008). Scores involve an age and education correction to determine whether performance is abnormal (Goldstein, 2000).

Although the development of the Luria-Nebraska battery was not without controversy (Adams, 1980; Spiers, 1981), it was an important landmark in neuropsychological assessment in that it provided a different fixed battery and introduced American neuropsychology to the ideas of Luria. It has been shown to discriminate between healthy and brain-damaged individuals, and compared with the HRNTB, it has been shown to be equally effective in identifying brain-damaged individuals (Tsushima, 2010). Nevertheless, as a comprehensive battery, it has not maintained the frequency of use over time, perhaps because of the psychometric limitations (Walker et al., 2008).

Although the HRNTB and the Luria-Nebraska battery were the first and most significant of the neuropsychological test batteries, other batteries have become increasingly popular in recent years. Two examples are the Neuropsychological Assessment Battery (Stern & White, 2003) and A Developmental Neuropsychological Assessment—the NEPSY (Korkman, 1988). The former is an updated and psychometrically sophisticated version of the batteries discussed earlier. The NEPSY is an outgrowth of Luria's approach for assessing children. These and other efforts indicate that there may be a resurgence of the battery approach in neuropsychological assessment.

Although the HRNTB and the Luria-Nebraska battery were vital for the development of clinical neuropsychology, as they provided evidence for neuropsychological evaluations as valuable tools for individuals with central nervous system dysfunction, the implementation of fixed batteries have declined among practicing neuropsychologists and the use of the flexible battery approach has increased (Rabin et al., 2005; Sweet et al., 2002). Collectively, the majority of neuropsychologists prefer a flexible battery approach (Sweet et al., 2002). The decline of the fixed battery approach and increase of the flexible approach may be related to the amount of time reimbursed by managed care, which calls for more a focused and time-sensitive approach, as is reflected in the flexible battery (Rabin et al., 2005).

Flexible battery. An alternative approach to the fixed battery was first proposed by Kaplan (Kaplan et al., 1978). This approach is considered more patient centered, as the battery of tests is selected based on the clinician's hypotheses to elucidate the patient's syndrome (Mitrushina, Boone, Razani, & D'Elia, 2005). The flexible battery approach allows practitioners to select measures to best understand

the patient's functioning, which is not possible in a fixed battery approach because clinicians cannot remove or add measures to the existing battery of tests. Given that neuropsychologists are able to target the problem with specific procedures and measures, it is suggested this approach is more time efficient and provides a more comprehensive understanding of the patient's difficulties (Bauer, 1999). The approach is influenced by a more European tradition in assessment, including Luria's approach, that does not have a specific set of tests or a rigid approach to understanding brain dysfunction.

The flexible battery approach is now favored by the majority of neuropsychologists, as it allows the professional to adjust and implement multiple measures and procedures to provide the most comprehensive understanding of the patient's difficulties (Bauer, 1999). Although the flexible battery approach is not without faults or restrictions, it indeed has become the most popular assessment approach among neuropsychologists (Rabin et al., 2005) and involves the administration of individual tests in different domains. Variability exists in the tests administered among practitioners for neuropsychological domains (e.g., Executive Functioning and Memory) with an interest in a flexible battery approach; however, there is typically commonality in neuropsychological domains assessed as well as tests administered.

Some writers, such as Faust (1991), have argued that the lack of standardization makes replication and acceptability incomplete in settings such as forensic ones. Because each case presents a unique situation and because each evaluation is customized to that situation, the underlying scientific support becomes eroded and its erosion poses problems in the legal arena. Reed (1996) outlined how the fixed battery—in this case, the HRNTB—was considered scientifically more rigorous than two flexible approaches. As a consequence, the flexible battery did not hold up to the scientific standards in legal situations, referred to as the *Daubert* standard. One possible way to address the variability of such an approach, at least with regard to the interpretation of the data, is to use a statistical method for interpretation outlined by Miller and Rohling (2001). Despite the current popularity of the flexible approach, the continued development of significant scientific underpinnings was encouraged 2 decades ago and has yet to be realized (Kane, 1991; Russell, Russell, & Hill, 2005). Regardless, the backbone of the flexible approach is a compendium of neuropsychological tests.

Most neuropsychological tests are grouped according to domain. Although differences exist, and there has yet to be an agreed-upon format, there is typically consistency among practitioners and approaches to determining the essential domains of a neuropsychological assessment. According to one of the most commonly used books in neuropsychological assessment, *Neuropsychological Assessment* (Lezak et al., 2004), the main domains are orientation and attention, perception, memory, verbal functions and language skills, construction, concept formation and reasoning, and executive and motor functions.

A problem that arises is that the categorization is variable, as different labels are used and categories with the same label have variable meanings. For example, sometimes the construct of "executive functioning" includes reasoning and problem solving, whereas in other situations it does not; sometimes attention is matched with orientation; other times, not. At present, there is no commonly accepted set of domains, or names and definitions of the domains, that neuropsychological assessment comprises. However, one empirical investigation using clinical psychologists and neuropsychologists from several professional organizations (e.g., APA and NAN) found the following common domains of tests: adaptive–functional, aphasia, behavioral medicine, developmental, intellectual or achievement, neuropsychological, and personality–psychopathology. Of these, intellectual and neuropsychological tests, followed by personality–psychopathology tests, were the most commonly used types of measures (Camara, Nathan, & Puente, 2000).

Finally, there are theoretical models such as the approach suggested by Luria in *The Working Brain* (1973) and *Higher Cortical Functions in Man* (1980). His model is based on an evolutionary and hierarchical system of behavior. Simpler behaviors, such as attention, are mediated by lower levels of the brain with more complex behaviors, such as executive functions, mediated by higher structures such

as the cerebral cortex. Domains are measured hierarchically, with the simpler or more fundamental behaviors measured first and more complex behaviors measured last. Thus, assessment of attention would precede executive functions, but if attention is impaired, the measurement of executive functions, in this case, would be fraught with error. Hence, assessing simpler functions may be necessary to make fundamental assumptions about more complex ones.

In a landmark book, Strauss et al. (2006) put together a compendium of neuropsychological tests. These authors reviewed and presented a large number of tests, allowing the neuropsychological community for the first time to have a comprehensive review of a larger number of individual instruments. Several studies have been published outlining what neuropsychological tests are used. One of the first articles on this topic appeared in 1987 when Peck outlined what he considered "essential" neuropsychological tests. The chapter indicated that it was not a survey, nor was it intended to be comprehensive.

Butler, Retzlaff, and Vanderploeg (1991) conducted one of the first comprehensive surveys through reviewing the *Journal of Clinical and Experimental Psychology*, *Neuropsychologia,* and the *International Journal of Clinical Neuropsychology* (no longer being published) between 1985 and 1989. A list of 116 neuropsychological tests was compiled, and the survey was mailed to 500 members of the International Neuropsychological Society (INS). In order of frequency, the following tests were reported as frequently being used: the WAIS (Wechsler, 1955), the Wechsler Memory Scale (WMS; Wechsler, 1945), the TMT (Reitan & Wolfson, 1985), the MMPI (Hathaway & McKinley, 1940), the Wide Range Achievement Test (Jastak & Jastak, 1965), the Bender–Gestalt test, portions of the HRNTB (Reitan & Wolfson, 1985), the Rorschach test (Exner, 1995), the Benton Visual Retention Test, the complete HRNTB (Reitan & Wolfson, 1985), the Wisconsin Card Sorting Test (Heaton, 1981), the Luria-Nebraska Neuropsychological Battery (Golden et al., 1978), and the Luria–Christenson Procedures. Finally, Butler et al. (1991) noted that the WAIS was used by 86% of the sample, and the next most frequently used measure, the WMS, was used half as often.

Camara et al. (2000) reported that the tests used by clinical neuropsychologists were not the same as those used by clinical psychologists (see Table 9.1). Overall, approximately 100 tests were frequently used, and most neuropsychologists used 25 tests very frequently and approximately another 25 "somewhat" frequently. The MMPI and the Wechsler scales, both intelligence and memory, were used by most of the sample. It is surprising that the MMPI was the most frequently used test by neuropsychologists.

At the 30th annual NAN conference in Vancouver, results from a national survey were presented (Ojeda & Puente, 2010). The study obtained a comprehensive list of neuropsychological tests that was based on a review of the literature, neuropsychological presentations, a review of the major test publishers, and a review of the *Buros Mental Measurements Yearbook*. A comprehensive list of 600 tests was obtained and sent to the members of the North Carolina Neuropsychological Society and the Pacific Northwest Neuropsychological Society as a limited sample test survey. Subsequently, an electronic list of these instruments was created and sent to members of Division 40 of APA and members of NAN. In order of prevalence, the most frequently used tests were as follows: the WAIS, the WMS, the TMT, the Wechsler Intelligence Scale for Children (WISC;

TABLE 9.1

Frequency of Tests Used by Clinical Neuropsychologists

Rank	Test
1	Minnesota Multiphasic Personality Inventory
2	Wechsler Adult Intelligence Scale—Revised
3	Wechsler Memory Scale—Revised
4	Trail-Making Test A and B
5	Finger Tapping Test
6	Grooved Pegboard Test
7	Hand Dynamometer
8	California Verbal Learning Test
9	Category Test
10	Wide Range Achievement Test—Revised and Third Editions

Wechsler, 1949), the Boston Naming Test (Kaplan et al., 1978), and the Rey-Oestereith Complex Figure Test (Rey, 1941). Although a total of 600 tests were reportedly being used, the frequency varied according to setting and not according to the geographical location.

Several specific test surveys have been conducted including for specific age groups, types of setting, and types of clients. In terms of specific age groups, Sellers (Sellars & Nadler, 1993) reported that the most frequently used tests for children were the WISC and the Wide Range Achievement Test—Revised (Jastak & Wilkinson, 1984). According to Sellers, the tests used with adults most frequently included portions of the HRNTB (e.g., the Category and Finger Tapping tests), the 1981 revised version of the WAIS, and the WMS. In terms of settings, the primary focus has been on determining whether neuropsychological test usage differs across clinical and forensic settings. Lees-Haley, Smith, Williams, and Dunn (1996) were the first to report that similar tests were used in both forensic and clinical settings, and Archer, Buffington-Vollum, Stredny, and Handel (2006) reported that frequently used neuropsychological tests in forensic setting were, in general, similar to those used in clinical ones. However, it appears that, although similar tests are used, the length of time involved in interpreting the tests is longer in the forensic setting.

Recent interest has arisen regarding the use of translated tests. Echemendia and Harris (2004) reported that similar tests were being used in English and Spanish and that the competency level of users varied considerably. Despite having access to almost 600 tests in Spanish, Ojeda and Puente (2010) reported that most neuropsychologists evaluating Spanish speakers used only approximately 50 of the available tests. Of those, approximately a dozen were frequently used. However, a NAN policy paper on testing Hispanics (Judd et al., 2009) warns about the simplistic translation and the use of North American norms for Spanish speakers. In Hong Kong, neuropsychological tests are used infrequently (Tsoi & Sundberg, 1989), and in China, Ryan, Dai, and Zheng (1994) collectively reported that the most frequently used tests in 1994 were the WAIS, the Chinese version of the WISC, the MMPI, the WMS, and the HRNTB.

A recent survey of 404 members of the NAN and the INS was conducted by Smith, Gorske, Wiggins, and Little (2010). The Beck Depression Scale (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) was the most commonly used test, followed by behavior ratings, and, subsequently, the Minnesota Multiphasic Personality Inventory—2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). In general, younger patients were more likely to have personality tests administered. Furthermore, personality tests were used most often when the evaluations did not involve classically brain-injured patients. Personality tests were more frequently used for learning disabilities and for forensic and clinical evaluations.

In recent years, neuropsychological testing has become more focused on the measurement of effort. Effort is broadly defined as the amount of motivation applied by the test taker. If motivation does not correlate well to test responses, the validity and reliability of the entire evaluation may come into question. Effort tests include, but are not limited to, the following: the b test (Boone et al., 2000), the Computerized Assessment of Response Bias (Lyell, Conder, & Green, 1997), the Dot Counting Test (Boone & Lu, 2002), the Test of Memory Malingering (Tombaugh, 1996), the Portland Digit Recognition Test (Binder, 2002), the Rey Memory Test (Reznek, 2005), the Victoria Symptom Validity Test (Slick, Hopp, & Straus, 1997), and the Word Memory Test (Green, 2005).

Norms

One of the most complicated aspects of neuropsychological tests is that of norms. Many test developers lack sufficient funds and personnel to mount significant standardization studies. Often small sample sizes are used, a problem compounded by the fact that many of the samples are geographically restricted and based on a limited clinical sample (e.g., dementia only). If obtained, "normals" (i.e., nonclinical samples) are sometimes not well matched to the clinical sample. Some tests (e.g., all versions of the Woodcock–Johnson) present primarily, if not exclusively, "normals." Other tests, such as the HRNTB, focus primarily on clinical samples. Other times, the norms are not well described, and the interpreter has to use a leap of faith in their

interpretation of the results. Finally, problems arise in that different norms exist. The Heaton norms, which are applicable to the HRNTB as well as to other commonly used tests, are the most frequently used, but even then they are limited by the sample size as well as other aspects (e.g., geographic limitations; Heaton, Grant, & Matthews, 1991). A related problem is whether norms from one ethnic group (e.g., Caucasian residents of the United States) could be used or are transferable to another, non-U.S. majority, ethnic group (e.g., Hispanics). Some, such as Ardila, Rosselli, and Puente (1992) have made limited norms with Spanish speakers available for some common tests such as the mini-mental status exam, but even these norms have problems. What ends up happening is that, regardless of the sensitivity of the test, its specificity ends up being affected by norms. It is not unusual that raw scores of one test may result in a normal interpretation with one set of norms and impaired with another set. In essence, the value of the norms is based on the referral question as well as the quality of the norms themselves. If one assumes that all norms are equal or valuable, it could result in errors in interpretation.

Technicians as Test Givers

Technicians are frequently involved with neuropsychological assessments to administer neuropsychological measures. In examining the Medicare utilization data, technicians are widely used in neuropsychological testing and infrequently used in psychological testing. Specifically, in the 2010 American Medical Association *Code Manager*, Medicare data reported indicated that psychological testing by a doctoral-level provider occurred 190,913 times, whereas psychological testing by a technician occurred in 13,009 instances. In contrast, neuropsychological testing by a doctoral-level provider occurred in 460,327 instances, whereas technicians provided the service 96,151 times.

Defining a technician is, ironically, both simple and difficult. According to the *Federal Register*, technicians are individuals who receive a 1099 form and, consequently, employees or independent contractors. They must hold a bachelor's degree from an accredited college or university with a major in an appropriate social or biological science (with at least 12 college credit hours in psychology). Furthermore, the federal government indicates that such individuals provide services under supervision. They typically administer and score tests but do not interpret tests or integrate test data with other sources of data prescribed by the supervisor. Additionally, they are suggested to have training in ethics, neuropsychology, psychopathology, and testing.

Specific to students, Medicare has never reimbursed for services provided by students in training for any health disciplines. The assumption is that general medical education pays training programs, and double dipping would occur if Medicare and the Current Procedural Terminology (CPT) reimbursed for student activity. However, students can perform as technicians as long as they are not being trained and their activity is not part of their educational requirements (e.g., a neuropsychologist in the community employs the student as a technician in his or her practice). Supervision can only be performed if the professional holds a doctoral degree in psychology, is licensed or certified as a psychologist, and is contractually related to the carrier that is being billed as a "clinical psychologist" (Centers for Medicare and Medicaid Services, 2004, p. 47553). On the plus side, technicians may increase the objectivity of data collection, minimize the potential for bias, and expand services available.

Time

Time is broadly defined as what the professional does while completing a neuropsychological evaluation. For neuropsychological testing, time is pretest, intratest, and posttest administration. *Pretest* is broadly defined as the time required for selecting and preparing the test. *Intratest* involves the actual administration of the test; *posttest* involves the scoring, interpretation, and integration of the test with other materials. This interpretation applies to both the neurobehavioral status exam (i.e., interview) as well as the testing done by the doctoral-level professional as well as the technician. For the technician, time that is billable is only face-to-face time (i.e., administration of the test). However, for purposes of payment for technician by the supervisor, time typically comprises test preparation, test administration,

and test scoring. Another way to determine time is to consider what it does not include: patient's completion of tests, scales, and forms; patient's waiting time; typing of reports; nonprofessional (e.g., clerical) time, and literature searches and learning new techniques.

Defining time specifically is based on "The Rounding Rule." According to the CPT, the following table would apply:

- 0 unit $<$ 31 minutes;
- 1 unit $\geq$ 31 minutes to $<$ 91 minutes;
- 2 units $\geq$ 91 minutes to $<$ 151 minutes;
- 3 units $\geq$ 151 minutes to $<$ 211 minutes;
- 4 units $\geq$ 271 minutes to $<$ 331 minutes, and so forth.

Another question is: How long is a neuropsychological battery of tests? The answer depends on the source of information. An examination of some of the previous studies reviewed (e.g., Sweet et al., 2002) found that the typical evaluation lasts well over 10 hours and, in some cases, upwards of over 15 hours. This reflects the earlier trends during the 1980s and 1990s, when evaluations were extremely lengthy, typically exceeding 10 and sometimes approaching 20 hours. Because of limitations imposed by managed care, the total amount of hours now typically do not exceed 10, largely because of industry caps on the total amount of time allocated. For forensic evaluations, however, these limits do not apply and may last up to 10 times longer than clinical assessments. In contrast, some evaluations are much shorter. For example, concussion evaluations onsite during sports activities (e.g., hockey and football) may last but a few minutes.

Also of importance is the ratio of time spent interviewing versus testing. In general, for every hour of interviewing, there are 5 hours of testing. According to Puente (2005), for every hour of test administration, a half hour of test scoring occurs, even though this varies considerably from test to test. For example, the TMT may take seconds to minutes to administer, whereas the Wechsler scales may take minutes to hours. Ball, Archer, and Imhof (1994) reported large differences for the 23 most commonly used tests based on administration, scoring, and time.

Interview and testing itself are typically reimbursable by insurance carriers, on average, for approximately 1 to 3 hours for interviewing and 6 to 10 hours for testing. Longer evaluations are often not reimbursed and may actually result in auditing by the insurance carrier. Typically, testing is part of direct patient contact and nondirect patient activity. The largest amount of time is the actual administration of tests as it consumes approximately two thirds of the total time. The final part includes scoring and interpretation of administered tests. For some tests, the scoring is easy and straightforward, but for others, the scoring can be laborious and time consuming. The most difficult portion is the interpretation or integration of test findings. In this portion, the qualified health professional integrates the results of the following sources of data for the final and integrated interpretation; record review, interview (direct and collateral), testing behavior, and test results. A written report provides a mechanism for documenting that the services were provided and, in turn, provides a method to communicate the information obtained to interested parties (e.g., referral source, patient, and collaterals).

Report

The standard written report contains several basic sections: Identifying Information, Reason for Evaluation, Evaluation Procedure, Tests and Testing Results, Integration, and Summary. A summary testing sheet providing specific numerical information sometimes accompanies the report as an appendix. Identifying Information contains data about the patient (age, gender, etc.). Reason for Evaluation identifies the referral source (e.g., neurologist) and purpose of the evaluation (e.g., assessment of memory). Evaluation Procedure explains what days (maybe even time of day) the work was performed and any aspect that would help replicate the study if one wanted to do so or if the case was to be audited. Tests and Testing Results vary as to whether a technician was involved or whether a professional did the entire testing. If a technician was involved, a section of specific information about the test administered is included as well as the actual results (e.g., number of errors on the Category test). The interpretation of that test in conjunction with other

information (e.g., history, other test results) is then included in an integrative fashion under a separate section, Integration and Summary as well as performed by the professional. If the professional does the testing, then Testing Results, Integration, and Summary can be placed under one section. The reason for the division of sections when a technician is used is to assist in understanding the report when an audit is being completed. Readers are encouraged to consult Chapter 3 in this volume, given the brevity of this section and the importance of report writing.

CHALLENGES FOR THE FUTURE OF NEUROPSYCHOLOGICAL ASSESSMENT

Clinical neuropsychological assessment has a long past but a short history. This history, however, has been explosive. With more than 100 years of application since Wundt, neuropsychology as a specialty started formally around 1980 with the organization of Division 40 of APA and NAN. Other organizations, such as the INS, have been more scientifically, rather than professionally, focused. However, it was not until 1996 that APA recognized clinical neuropsychology as a specialty (APA Commission for the Recognition of Specialties and Proficiencies in Professional Psychology, 1996). During those years and since then, growth has been dramatic. Neuropsychology has grown to be the primary clinical assessment in psychology and the largest group of diagnosticians as well as clinical testers in psychology. With this growth, there has also been a drastic increase in tests and patterns of testing. There are probably well over 2,000 tests currently being used in clinical neuropsychology although probably only 50 to 100 are used with some regularity. The specialty has gone from relying almost exclusively on batteries (e.g., the HRNTB) to almost exclusively a composite of individual tests (e.g., the WAIS). Finally, neuropsychological testing has gone from being a strictly clinical enterprise, focusing initially on neurological and psychiatric patients, to addressing varied populations (e.g., sports, military, and health) as well as forensic ones. It appears, however, that this enormous growth may not be as strongly supported scientifically as it should, and its application to ethnic minority groups (e.g., Hispanics) remains relatively weak.

Three challenges lie ahead for neuropsychological assessment: (a) The scientific base needs to be expanded, and translational research needs to be a primary focus; (b) there needs to be an understanding of nonmajority group members, especially in light of shifting American demographics and the globalization of neuropsychological assessment increases; and (c) inclusion of neuropsychological assessment in wide-spectrum health and related fields (e.g., education, sports, law) needs to occur.

In addition to these three challenges, historical problems persist, including being perceived as overly political, inbred, and elitist. This perception, real or otherwise, may impede the generalizability of neuropsychological assessment to wider audiences, both geographically (e.g., to developing countries) and for other specialties within psychology (e.g., industrial psychology). These problems may prevent Wundt and James's beliefs about psychology from being heavily associated with underlying brain function and may limit the role that neuropsychological approaches play in answering traditional philosophical questions. Regardless, the explosive growth of clinical neuropsychology and neuropsychological assessment over the past 3 decades potentially signals a paradigm shift within the measurement of abnormal behavior.

References

Adams, K. M. (1980). In search of Luria's battery: A false start. *Journal of Consulting and Clinical Psychology, 48*, 511–516. doi:10.1037/0022-006X.48.4.511

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association Division 40 Executive Committee. (2006). *Definition.* Retrieved from http://www.div40.org/pub/archival_definition.html

Archer, R. P., Buffington-Vollum, J. K., Stredny, R., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94. doi:10.1207/s15327752jpa8701_07

Ardila, A., Rosselli, M., & Puente, A. E. (1992). *Neuropsychological evaluation of the Spanish speaker*. New York, NY: Plenum Press.

Ball, J. D., Archer, R. P., & Imhof, E. A. (1994). Time requirements of psychological testing: A survey of practitioners. *Journal of Personality Assessment, 63*, 239–249. doi:10.1207/s15327752jpa6302_4

Barth, J. T., Pliskin, N., Axelrod, B., Faust, D., Fisher, J., Harley, J. P., . . . Silver, C. (2003). Introduction to the NAN 2001 Definition of a Clinical Neuropsychologist: NAN Policy and Planning Committee. *Archives of Clinical Neuropsychology*, *18*, 551–555.

Bauer, R. M. (1999). The flexible battery approach to neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 419–448). Hillsdale, NJ: Erlbaum.

Beaumont, G. (2008). *Introduction to neuropsychology*. New York, NY: Guilford Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571. doi:10.1001/archpsyc.1961.01710120031004

Benton, A. L. (1972). Psychological tests for brain damage. In A. M. Freedman & H. I. Kaplan (Eds.), *Diagnosing mental illness: Evaluation in psychiatry and psychology* (pp. 62–80). Oxford, England: Atheneum.

Benton, A. L. (1992). Clinical neuropsychology: 1960–1990. *Journal of Clinical and Experimental Neuropsychology, 14*, 407–417. doi:10.1080/01688639208407616

Binder, L. (2002). The Portland digit recognition test. *Journal of Forensic Neuropsychology, 2*, 27–41. doi:10.1300/J151v02n03_02

Boake, C. (2008). Clinical neuropsychology. *Professional Psychology: Research and Practice, 39*, 234–239. doi:10.1037/0735-7028.39.2.234

Boone, K., & Lu, P. (2002). *The dot counting test*. Los Angeles, CA: Western Psychological Services.

Boone, K. B., Lu, P., Sherman, D., Palmer, B., Back, C., Shamieh, E., . . . Berman, N. (2000). Validation of a new technique to detect malingering of cognitive symptoms. *Archives of Clinical Neuropsychology*, *15*, 227–241.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory—2 (MMPI–2): Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butler, M., Retzlaff, P. D., & Vanderploeg, R. (1991). Neuropsychological test usage. *Professional Psychology: Research and Practice, 22*, 510–512. doi:10.1037/0735-7028.22.6.510

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi:10.1037/0735-7028.31.2.141

Centers for Medicare and Medicaid Services. (2004). Medicare program: Revisions to payment policies under the physician fee schedule for calendar year 2005. *Federal Register*, *69*(150), 47488–47730.

Demakis, G. J. (2004). Frontal lobe damage and tests of executive processing: A meta-analysis of the Category test, Stroop test, and Trail-Making Test. *Journal of Clinical and Experimental Neuropsychology, 26*, 441–450. doi:10.1080/13803390490510149

Echemendia, R. J., & Harris, J. G. (2004). Neuropsychological test use with Hispanic/Latino populations in the United States: Part II of a national survey. *Applied Neuropsychology, 11*, 4–12. doi:10.1207/s15324826an1101_2

Elias, M. F., Elias, P. K., Sullivan, L. M., Wolf, P. A., & D'Agostino, R. B. (2003). Lower cognitive function in the presence of obesity and hypertension: The Framingham Heart Study. *International Journal of Obesity, 27*, 260–268. doi:10.1038/sj.ijo.802225

Exner, J. E., Jr. (1995). *A Rorschach workbook for the comprehensive system* (4th ed.). Asheville, NC: Rorschach Workshops.

Faust, D. (1991). Forensic neuropsychology: The art of practicing a science that does not yet exist. *Neuropsychology Review, 2*, 205–231. doi:10.1007/BF01109045

Fennell, E. B. (2000). Issues in child neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 357–381). Hillsdale, NJ: Erlbaum.

Fitzhugh-Bell, K. B. (1997). Historical antecedents of clinical neuropsychology. In A. M. Horton, D. Wedding, & J. Webster (Eds.), *The neuropsychology handbook: Vol. 1. Foundations and assessment* (2nd ed., pp. 67–90). New York, NY: Springer.

Franz, S. (1920). *Handbook of mental examination methods* (2nd ed.). New York, NY: Macmillan.

Golden, C. J. (1982). *Item interpretation of the Luria-Nebraska Neuropsychological Battery*. Lincoln: University of Nebraska Press.

Golden, C. J., Hammeke, T. A., & Purisch, A. D. (1978). Diagnostic validity of a standardized neuropsychological battery derived from Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology, 46*, 1258–1265.

Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1985). *Luria-Nebraska Neuropsychological Battery: Forms I and II manual*. Los Angeles, CA: Western Psychological Services.

Goldstein, G. (1985). The history of clinical neuropsychology: The role of some American pioneers. *International Journal of Neuroscience, 25*, 273–275. doi:10.3109/00207458508985380

Goldstein, G. (2000). Comprehensive neuropsychological assessment batteries. In G. Goldstein & M. Hersen

(Eds.), *Handbook of psychological assessment* (pp. 231–262). Kidlington, England: Elsevier.

Goldstein, G., Incagnoli, T., & Puente, A. E. (2011). *Contemporary neuropsychological syndromes.* New York, NY: Springer.

Goldstein, K. (1942). *Aftereffects of brain injuries in war: Their evaluation and treatment; the application of psychologic methods in the clinic.* New York, NY: Grune & Stratton.

Green, P. (2005). *Word memory test for Microsoft Windows: User's manual.* Edmonton, Alberta, Canada: Green Publications.

Hartlage, L. C., & Long, C. J. (2009). Development of neuropsychology as a professional psychological specialty: History, training, and credentialing. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (3rd ed., pp. 3–18). New York, NY: Springer Science + Business Media.

Hartman, D. E. (1991). Reply to Reitan: Unexamined premises and the evolution of clinical neuropsychology. *Archives of Clinical Neuropsychology, 6*, 147–165.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule: I. Construction of the schedule. *Journal of Psychology, 10*, 249–254. doi:10.1080/00223980.1940.9917000

Heaton, R. K. (1981). *Wisconsin Card Sorting Test manual.* Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead–Reitan Battery*. Odessa, FL: Psychological Assessment Resources.

Horton, A. M. (2008). The Halstead–Reitan Neuropsychological Test Battery: Past, present, and future. In A. Horton & D. Wedding (Eds.), *The neuropsychology handbook* (pp. 251–278). New York, NY: Springer.

Horton, A. M. (2010). Overview of forensic neuropsychology. In A. M. Horton & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 3–9). New York, NY: Springer.

Howieson, D. B., & Lezak, M. D. (2010). The neuropsychological evaluation. In S. C. Yudofsky & R. E. Hales (Eds.), *Essentials of neuropsychiatry and behavioral neurosciences* (2nd ed., pp. 29–54). Arlington, VA: American Psychiatric Press.

James, W. (1890). *Principles of psychology*. Cambridge, MA: Harvard University Press. doi:10.1037/11059-000

Jastak, J. F., & Jastak, S. R. (1965). *The Wide Range Achievement Test.* Wilmington, DE: Guidance Associates.

Jastak, S., & Wilkinson, G. S. (1984). *Wide Range Achievement Test—Revised.* Wilmington, DE: Jastak Associates.

Johnson, J. A., & D'Amato, R. C. (2011). Examining and using the Halstead-Reitan Neuropsychological Test Battery: Is it our future or our past? In A. S. Davis (Ed.), *Handbook of pediatric neuropsychology* (pp. 353–366). New York, NY: Springer.

Judd, T. T., Capetillo, D., Carrion-Baralt, J., Marmol, L. M., San Miguel-Montez, L., Navarratte, M. G., . . . Valdez, J. (2009). Professional consideration for improving the neuropsychological evaluation of Hispanics: A National Academy of Neuropsychology education paper. *Archives of Clinical Neuropsychology, 24*, 127–135. doi:10.1093/arclin/acp016

Kane, R. L. (1991). Standardized and flexible batteries in neuropsychology: An assessment update. *Neuropsychology Review, 2*, 281–339. doi:10.1007/BF01108849

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1978). *The Boston Naming Test: Experimental edition.* Philadelphia, PA: Lea & Febiger.

Korkman, M. (1988). NEPSY—An adaptation of Luria's investigation for young children. *Clinical Neuropsychologist, 2*, 375–392. doi:10.1080/13854048808403275

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology, 11*, 45–51. doi:10.1016/0887-6177(95)00011-9

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York, NY: Oxford University Press.

Libon, D. J., Xie, S. X., Moore, P., Farmer, J., Antani, S., McCawley, G., . . . Grossman, M. (2007). Patterns of neuropsychological impairment in frontotemporal dementia. *Neurology*, *68*, 369–375. doi:10.1212/01.wnl.0000252820.81313.9b

Luria, A. (1973). *Working brain.* London, England: Penguin Books.

Luria, A. (1980). *Higher cortical functions in man.* New York, NY: Basic Books. doi:10.1007/978-1-4615-8579-4

Lyell, L., Conder, B., & Green, M. (1997). *Computerized assessment of response bias.* Durham, NC: CogniSyst.

Marcotte, T. D., Scott, J. C., Kamat, R., & Heaton, R. (2010). Neuropsychology and the prediction of everyday functioning. In T. D. Marcotte & I. Grant (Eds.), *Neuropsychology of everyday functioning* (pp. 5–38). New York, NY: Guilford Press.

McCaffrey, R. J., Williams, A. D., Fisher, J. M., & Laing, L. C. (1997). *The practice of forensic neuropsychology: Meeting challenges in the courtroom.* New York, NY: Plenum Press.

McClintock, S. M., Husain, M. M., Greer, T. L., & Cullum, C. (2010). Association between depression

severity and neurocognitive function in major depressive disorder: A review and synthesis. *Neuropsychology, 24*, 9–34. doi:10.1037/a0017336

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease. Report of the NINCDS–ADRDA work group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology, 34*, 939–944. doi:10.1212/WNL.34.7.939

Meier, M. J. (1997). The establishment of clinical neuropsychology as a specialty. In M. E. Maruish & J. A. Moses (Eds.), *Clinical neuropsychology: Theoretical foundations for practitioners* (pp. 1–32). Mahwah, NJ: Erlbaum.

Miller, L. S., & Rohling, M. (2001). A statistical interpretive method for neuropsychological test data. *Neuropsychology Review, 11*, 143–169. doi:10.1023/A:1016602708066

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Ojeda, C., & Puente, A. E. (2010). *Neuropsychological testing of Spanish speakers: The challenge of accurately assessing linguistically and culturally diverse individuals* (Unpublished master's thesis). University of North Carolina, Wilmington.

Ostrosky-Solís, F., Ardila, A., & Rosselli, M. (1999). Neuropsi: A brief neuropsychological test battery in Spanish with norms by age and educational level. *Journal of the International Neuropsychological Society, 5*, 413–433. doi:10.1017/S1355617799555045

Palmer, B. W., Dawes, S. E., & Heaton, R. K. (2009). What do we know about neuropsychological aspects of schizophrenia? *Neuropsychology Review, 19*, 365–384. doi:10.1007/s11065-009-9109-y

Parsons, O. A., & Butters, N. E. (1987). *Neuropsychology of alcoholism: Implications for diagnosis and treatment*. New York, NY: Guilford Press.

Porter, R. J., Bourke, C., & Gallagher, P. (2007). Neuropsychological impairment in major depression: Its nature, origin and clinical significance. *Australian and New Zealand Journal of Psychiatry, 41*, 115–128. doi:10.1080/00048670601109881

Prichard, D. (1997). Forensic neuropsychology. In M. E. Maruish & J. A. Moses (Eds.), *Clinical neuropsychology: Theoretical foundations for practitioners* (pp. 81–118). Mahwah, NJ: Erlbaum.

Puente, A. E. (1989). Historical perspectives in the development of neuropsychology as a professional psychological specialty. In C. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of child clinical neuropsychology* (pp. 3–16). New York, NY: Plenum Press.

Puente, A. E. (2005). Some lessons I have learned from 25 years in clinical neuropsychology: A letter to my grandchildren. *Neuropsychology Review, 15*, 197–207. doi:10.1007/s11065-005-9181-x

Puente, A. E., Adams, R., Barr, W. B., Bush, S. S., Ruff, R. M., Barth, J. T., . . . Tröster, A. I. (2006). The use, education, training and supervision of neuropsychological test technicians (psychometrists) in clinical practice. Official Statement of the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, *21*, 837–839. doi:10.1016/j.acn.2006.08.011

Puente, A. E., & Marcotte, A. C. (2000). A history of Division 40 (Clinical Neuropsychology). In D. A. Dewsbury (Ed.), *Unification through division: Histories of the divisions of the American Psychological Association* (pp. 137–160). Washington, DC: American Psychological Association. doi:10.1037/10356-006

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33–65. doi:10.1016/j.acn.2004.02.005

Reed, J. E. (1996). Fixed vs. flexible neuropsychological test under the Daubert standard for the admissibility of scientific evidence. *Behavioral Sciences and the Law, 14*, 315–322. doi:10.1002/(SICI)1099-0798(199622)14:3<315::AID-BSL242>3.0.CO;2-X

Reitan, R. M. (1989). A note regarding some aspects of the history of clinical neuropsychology. *Archives of Clinical Neuropsychology, 4*, 385–391.

Reitan, R. M. (1994). Ward Halstead's contributions to neuropsychology and the Halstead-Reitan Neuropsychological Test Battery. *Journal of Clinical Psychology, 50*, 47–70. doi:10.1002/1097-4679(199401)50:1<47::AID-JCLP2270500106>3.0.CO;2-X

Reitan, R. M., & Davison, L. A. (Eds.). (1974). *Clinical neuropsychology: Current status and applications.* Oxford, England: V. H. Winston & Sons.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead–Reitan Neuropsycholgical Test Battery: Therapy and clinical interpretation.* Tucson, AZ: Neuropsychological Press.

Reitan, R. M., & Wolfson, D. (2004). Theoretical, methodological, and validational bases of the Halstead–Reitan Neuropsychological Test Battery. In G. Goldstein, S. R. Beers, & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment: Vol. 1. Intellectual and neuropsychological assessment* (pp. 105–131). Hoboken, NJ: Wiley.

Rey, A. (1941). L'Examen psychologique dans les cas d'encephalopathie traumatique [The psychological

examination in cases of traumatic encephalopathy]. *Archives de Psychologie, 28*, 215–285.

Reynolds, C. R., Castillo, C. L., & Horton, A. M. (2008). Neuropsychology and intelligence. In A. Horton & D. Wedding (Eds.), *The neuropsychology handbook* (pp. 70–86). New York, NY: Springer.

Reznek, L. (2005). The Rey 15-item memory test for malingering: A meta-analysis. *Brain Injury, 19*, 539–543. doi:10.1080/02699050400005242

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structural interview of reported symptoms.* Odessa, FL: Psychological Assessment Resources.

Russell, E. W., Russell, S. L. K., & Hill, B. D. (2005). The fundamental psychometric status of neuropsychological batteries. *Archives of Clinical Neuropsychology, 20*, 785–794. doi:10.1016/j.acn.2005.05.001

Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach.* Oxford, England: Wiley-Interscience.

Ryan, J. J., Dai, X., & Zheng, L. (1994). Psychological test usage in the People's Republic of China. *Journal of Psychoeducational Assessment, 12*, 324–330. doi:10.1177/073428299401200402

Sbordone, R. J., & Saul, R. E. (2000). *Neuropsychology for health care professionals and attorneys* (2nd ed.). Boca Raton, FL: CRC Press.

Sellers, A. H., & Nadler, J. D. (1993). A survey of current neuropsychological assessment procedures used for different age groups. *Psychotherapy in Private Practice*, *11*, 47–57. doi:10.1300/J294v11n03_10

Skinner, B. F. (1953). *Science and human behavior.* Oxford, England: Macmillan.

Slick, D. J., Hopp, G., & Straus, E. (1997). *Victoria symptom validity test.* Odessa, FL: Psychological Assessment Resources.

Smith, S. R., Gorske, T. T., Wiggins, C., & Little, J. A. (2010). Personality assessment use by clinical neuropsychologists. *International Journal of Testing, 10*, 6–20. doi:10.1080/15305050903534787

Sperry, R. W. (1995). The future of psychology. *American Psychologist, 50*, 505–506. doi:10.1037/0003-066X.50.7.505

Spiers, P. A. (1981). Have they come to praise Luria or to bury him? The Luria-Nebraska Battery controversy. *Journal of Consulting and Clinical Psychology, 49*, 331–341. doi:10.1037/0022-006X.49.3.331

Stern, R. A., & White, T. (2003). *Neuropsychological Assessment Battery.* Odessa, FL: Psychological Assessment Resources.

Storey, E., Slavin, M. J., & Kinsella, G. J. (2002). Patterns of cognitive impairment in Alzheimer's disease: Assessment and differential diagnosis. *Frontiers in Bioscience, 7*, e155–e184. doi:10.2741/storey

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford University Press.

Sweet, J. J., Peck, E., Abramowitz, C., & Etzweiler, S. (2002). National Academy of Neuropsychology/Division 40 of the American Psychological Association Practice survey of clinical neuropsychology in the United States, Part I: Practitioner and practice characteristics, professional activities, and time requirements. *The Clinical Neuropsychologist, 16*, 109–127. doi:10.1076/clin.16.2.109.13237

Tombaugh, T. (1996). *Test of memory malingering.* New York, NY: Multi-Health Systems.

Tonkonogy, J., & Puente, A. E. (2009). *Localization of clinical syndromes in neuropsychology and neuroscience.* New York, NY: Springer.

Tsoi, M. M., & Sundberg, N. D. (1989). Patterns of psychological test use in Hong Kong. *Professional Psychology: Research and Practice, 20*, 248–250. doi:10.1037/0735-7028.20.4.248

Tsushima, W. T. (2010). Luria-Nebraska neuropsychological battery. In I. B. Weiner & E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., pp. 950–952). Hoboken, NJ: Wiley. doi:10.1002/9780470479216.corpsy0519

Tully, P. J., Baker, R. A., Knight, J. L., Turnbull, D. A., & Winefield, H. R. (2009). Neuropsychological function 5 years after cardiac surgery and the effect of psychological distress. *Archives of Clinical Neuropsychology, 24*, 741–751. doi:10.1093/arclin/acp082

Vanderploeg, R. D. (2000). Interview and testing: The data collection phase of neuropsychological evaluations. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed., pp. 3–36). Mahwah, NJ: Erlbaum.

Vardy, J., Rourke, S., & Tannock, I. F. (2007). Evaluation of cognitive function associated with chemotherapy: A review of published studies and recommendations for future research. *Journal of Clinical Oncology, 25*, 2455–2463. doi:10.1200/JCO.2006.08.1604

Walker, J., D'Amato, R. C., & Davis, A. (2008). Understanding and using the Luria-Nebraska Neuropsychological Test Batteries with children and adults. In R. C. D'Amato & L. C. Hartlage (Eds.), *Essentials of neuropsychological assessment: Rehabilitation planning for intervention* (2nd ed., pp. 127–148). New York, NY: Springer.

Watson, J. (1919). *Psychology from the standpoint of a behaviorist.* Philadelphia: Lippincott. doi:10.1037/10016-000

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology, 19*, 87–95. doi:10.1080/00223980.1945.9917223

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children.* Oxford, England: Psychological Corporation.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* New York, NY: Psychological Corporation.

Wundt, W. (1904). *Principles of physiological psychology* (E. B. Titchener, Trans.). London, England: Swan Sonnenschein.

Yochim, B. (2010). Consideration of neuropsychological factors in interviewing. In D. L. Segal & M. Hersen (Eds.), *Diagnostic interviewing* (pp. 103–124). New York, NY: Springer. doi:10.1007/978-1-4419-1320-3_6

Zillmer, E. A. (2004). National Academy of Neuropsychology: President's address: The future of neuropsychology. *Archives of Clinical Neuropsychology, 19*, 713–724. doi:10.1016/j.acn.2004.05.004

CHAPTER 10

ASSESSMENT OF PERSONALITY AND PSYCHOPATHOLOGY WITH PERFORMANCE-BASED MEASURES

Irving B. Weiner

Psychological tests for assessing personality functioning and psychopathology are of two kinds: self-report inventories (SRIs) and performance-based measures (PBMs). Of these, SRIs are the more familiar and widely used types of tests. Typically, they consist of self-descriptive statements (e.g., "I often lose my temper") that are rated for their accuracy by answering "yes" or "no" or by selecting a point on a Likert-type scale (e.g., alternatives ranging from "very much like me" to "not at all like me," with numerical scale anchors). SRIs are relatively structured tests that inquire fairly directly about a person's self-perceptions, state of mind, and life experiences. They can be computer scored and quantified, but they are not completely objective in nature. In the absence of specific benchmarks, SRI items may require subjective judgments. For example, how respondents rate statements like "I often lose my temper" depends on how they define the frequency of "often" and what they consider to constitute losing one's temper. Additional discussion of SRIs can be found in Chapter 11 of this volume, which is devoted to this topic.

In contrast to SRIs, PBMs sample not what people say about themselves but how they perform on various tasks. For example, if asked to copy a design, do they do so carefully, neatly, and with precision, or do they draw carelessly, sloppily, and with little regard for accuracy? As an alternative to asking people directly whether they regard themselves as careful, neat, and precise, these qualities are inferred indirectly from the manner in which they draw the design. This PBM type of indirect assessment is less structured than SRI assessment and necessitates examiner coding of responses instead of automatic computer scoring of checkmarks on an answer sheet. The codes that examiners assign to features of PBM responses can nevertheless be quantified to facilitate their interpretation and assessment of their psychometric properties.

Just as SRIs are not entirely objective, PBM assessment is not an entirely subjective process, provided that adequate coding guidelines are available. For example, if people are asked what an inkblot might look like and respond, "Looking at all of it, it could be a bat," the response can be coded *W* for Whole, and using the whole blot and coding a response as *W* are objective matters without error variance. Additionally, although the relatively unstructured PBMs are sometimes referred to as free-response measures, they typically include some unambiguous stimulus elements and some specific instructions for what respondents are expected to do.

In this latter regard, it should be noted that PBMs were, for many years, referred to as *projective* tests, following a suggestion by Frank (1939) that relatively unstructured personality tests induce a respondent to "project upon that plastic field . . . his private world of personal meanings and feelings" (pp. 395, 402). Few would question that personality characteristics are likely to influence how people interpret ambiguous stimulus fields. In recent years, however, assessment psychologists have come to recognize increasingly that many features of PBM responses are produced and can be interpreted independently of projective processes. To take as an

DOI: 10.1037/14048-010
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

example the previously mentioned inkblot response based on the entire configuration of the blots, it can be said that these *W* responses identify a global focus of attention; that a high percentage of *W* responses probably signifies an inclination to view events in global terms, perhaps with insufficient attention to details; and that both this response style and this inference from it are a matter of perceptual–cognitive processes that do not involve projection. PBM is thus a broader and more accurate way of describing indirect personality assessment measures than *projective*, and the misleading connotations of traditional distinctions between *objective* and *projective* personality tests have been recognized in recent publication guidelines like the following: "The *Journal of Personality Assessment* will facilitate the transition to more adequately differentiated assessment terminology by asking authors to avoid referring to categories of personality tests as objective or projective" (Meyer & Kurtz, 2006, pp. 224–225).

This chapter on PBM assessment describes how these measures contribute to the personality assessment process, how they identify state and trait characteristics of individuals, and how they assist differential diagnosis of psychological disorder. After these descriptions, the text reviews the format, history, psychometric foundations, and applications of the two most widely used personality assessment PBMs: the Rorschach Inkblot Method (RIM) and the Thematic Apperception Test (TAT).

CONTRIBUTION OF PERFORMANCE-BASED MEASURES TO THE PERSONALITY ASSESSMENT PROCESS

The distinctive features of PBMs contribute to the personality assessment process in three ways. As indirect measures, first of all, PBMs are more likely than SRIs to elicit clues to personal characteristics and behavioral tendencies of which individuals are not fully aware or are reluctant to divulge. The relatively direct SRIs have the advantage of being less inferential than PBMs, but the information they obtain is limited to what people are able and willing to say about themselves. With regard to what they are able to say about themselves, most people have feelings and attitudes that are outside their conscious awareness, and most people have had life experiences that they recall vaguely, inaccurately, or not at all. As for what they are willing to say, most people know things about themselves that they would rather not disclose, whether because of embarrassment or because of possible unwelcome consequences of the disclosure. Hence, although PBM interpretations often must be more inferential and less definitive than conclusions based on SRI responses, the indirectness of PBMs can result in clues to personality characteristics that might not become evident in assessments limited to SRIs.

The second contribution of PBMs to the personality assessment process flows from their minimal face validity, which tends to make them less susceptible than SRIs to impression management. The directness of many SRI items gives them an extent of face validity that makes their implications apparent. For example, responding "true" to "I cannot do anything well" or "I cry easily" is clearly a statement of low self-esteem in the first instance and of being overly emotional in the second instance. By contrast, the indirect nature of PBM responses, as in saying what an inkblot might look like, gives respondents little inkling of what their responses might signify. Hence, test takers intent on appearing more psychologically sound or more emotionally upset than is actually the case usually have more difficulty tailoring their PBM responses than their SRI responses to achieve these effects (see Bornstein, Rossner, Hill, & Stepanian, 1994; Weiner, 2005a).

Third, because PBM personality assessments customarily involve a verbal exchange between examiner and examinee, they provide more occasions for interpersonal interaction than the typical self-administration of SRIs. PBM testing is, therefore, likely to generate more information than SRI testing about such matters as a respondent's interpersonal style (e.g., assertive or deferential), test-taking attitudes (e.g., cooperative or resistive), and expectations of success or failure when facing challenging tasks (e.g., optimistic and self-confident or defeatist and self-denigrating). Although these styles, attitudes, and expectations often can be identified from item endorsements on self-administered SRIs, PBM testing gives examiners an opportunity to observe firsthand how examinees act and what they have to say.

With appreciation for the relative advantages and limitations of both PBMs and SRIs, assessment psychologists generally concur that personality test batteries should include both kinds of instruments. Personality assessments that integrate data from different types of tests will almost always paint a fuller and more accurate picture of an individual's personality functioning than testing that consists only of PBMs or SRIs (see Beutler & Groth-Marnat, 2003; Weiner, 2005a). This complementarity of PBMs and SRIs reflects recommendations made many years ago by Campbell and Fiske (1959) for multimethod assessment in psychology, and the Psychological Assessment Work Group, a task force appointed by the American Psychological Association Board of Professional Affairs to examine the utility of psychological assessment methods, has similarly concluded that conjoint testing with self-report and performance-based instruments is how "practitioners have historically used the most efficient means at their disposal to maximize the validity of their judgments about individual clients" (Meyer et al., 2001, p. 150).

PERFORMANCE-BASED ASSESSMENT OF PERSONALITY AND PSYCHOPATHOLOGY

PBMs contribute to the assessment of personality and psychopathology by providing information about personality characteristics and assisting in the differential diagnosis of psychological disorder. Sources of data in PBM assessments can identify a broad range of current states and personality traits, and indications of particular states and traits can help determine the presence and severity of psychological disorders that are defined at least in part by these characteristics.

PROVIDING INFORMATION ABOUT PERSONALITY CHARACTERISTICS

Personality characteristics consist of an individual's states and traits. *States* refer to the thoughts and feelings a person is having at the moment, as in being attentive or distracted, anxious or unperturbed, happy or sad, bored or enthusiastic. *Traits* refer to an individual's abiding dispositions to think, feel, or act in certain ways and to deal with similar situations in a similar manner, as in being generally a kind, irresponsible, thrifty, or abrasive type of person. PBMs identify states and traits by examining the ways in which people tend to look at and perceive situations (e.g., globally or narrowly, accurately or mistakenly); how they conceptualize and draw conclusions about relationships between events (e.g., cautiously or hastily, sensibly or illogically); the nature and adequacy of their preferred means of dealing with stressful situations (effective or ineffectual, active struggle or passive withdrawal); how they experience and express their emotions (e.g., intensely or with little depth, dramatically or with reserve); and their attitudes toward themselves, other people, and social interactions (e.g., positive or negative, caring or indifferent).

Examiners glean information about these manifestations of state and trait characteristics from three sources of data in PBMs: (a) structural features of test responses, (b) imagery and associations evoked by the test stimuli, and (c) behavioral observations of how respondents conduct themselves during an examination and relate to the examiner. The structural features of PBMs consist of objective aspects of test responses that are representative samples of the behavioral tendencies typically inferred from them. As previously mentioned in the inkblot example, a high percentage of *W* (Whole) responses in a Rorschach test protocol usually indicates a tendency to perceive situations in global terms while giving insufficient attention to details that should also be noticed.

The imagery and evoked associations in PBMs constitute imaginal productions that express or symbolize a person's attitudes, concerns, and behavioral tendencies. On the TAT, for example, in which people are asked to make up stories about a series of pictures, stories in which the central figures are enjoying success or have been victimized in some way suggest a high need for achievement in the first instance and feelings of vulnerability to being harmed or exploited in the second instance.

The behavioral data in PBM assessment include the manner in which test takers approach the tasks set for them, whatever attitudes they express toward themselves or toward the examiner, how they speak and use language, and their facial expressions and

bodily posturing. The careful or careless rendering of drawings and an assertive or deferential interpersonal style in relating to the examiner, both mentioned previously, illustrate such behavioral data. Complimentary or depreciatory comments about their test performance may reflect positive or negative self-attitudes of respondents (e.g., "That was a pretty interesting story, wasn't it?" vs. "I'm not any good at making up stories"). Respondents may speak fluently or haltingly, use formal or casual language, and talk grammatically with a rich vocabulary or ungrammatically with a limited one, each of which may be a representative sample of their verbal style in other interpersonal situations. As for facial expressions and bodily posturing, how examinees feel at the moment may be suggested by such manifestations as profuse sweating (anxious), a clenched jaw (angry), a mournful expression (sad), or constant smiling or laughing (euphoric or sometimes with a forced gaiety that suggests efforts to ward off underlying depression).

ASSISTING DIFFERENTIAL DIAGNOSIS OF PSYCHOLOGICAL DISORDER

PBMs, in common with SRIs, indicate psychologically healthy functioning when the test results fall within a normally expected range. Results for a test variable that deviate markedly from normal expectation are likely to indicate maladaptive dysfunction in the aspect of personality functioning measured by this variable. However, this guideline must be applied with two considerations in mind. First, maladaptive dysfunction is culturally relative in the sense that a personality characteristic that is deviant in one culture may not be deviant in another culture. There is good reason to believe that personality characteristics are a universal phenomenon and that personality assessment methods can be used effectively to identify what people are like wherever they are (see Dana, 2005; McCrae & Terracciano, 2008). As a matter of cultural relativism, however, the implications of an individual's personality characteristics for his or her psychological adjustment are not universal; they depend instead on the person's cultural context, consisting of whatever national, ethnic, religious, neighborhood, family, or other group values play an influential part in his or her life.

For example, an independent-minded person is more likely to feel good and fare well in a cultural context that encourages individualism and admires self-reliance than in surroundings that emphasize group decision making and reward conformity. Conversely, a person with a dependent nature would have better prospects of adapting successfully in a group-oriented rather than in an individualistic cultural setting. The use of personality descriptions as a basis for estimating how people will respond to and deal with events in their lives requires keen awareness of the culturally based values and expectations that have a bearing on these events.

The second consideration to keep in mind in evaluating psychological disorder is that PBMs do not assess psychopathology directly, at least not with respect to the categories of disorder in the *Diagnostic and Statistical Manual of Mental Disorders* (4th edition, text revision; *DSM–IV–TR*; American Psychiatric Association, 2000). *DSM* diagnostic criteria consist largely of manifest symptoms or behavioral patterns that are not directly measured by PBMs (e.g., delusions for schizophrenia, insomnia for major depressive disorder [MDD], recurrent nightmares for posttraumatic stress disorder [PTSD], irresponsibility for antisocial personality disorder [ASP], and social or occupational dysfunction for numerous disorders). Moreover, many *DSM* criteria specify the duration of symptoms or behavior patterns (e.g., at least 6 months for schizophrenia; at least 2 weeks for MDD; more than 1 month for PTSD; since childhood for ASP) that are not state or trait characteristics of an individual and cannot be determined from PBM data.

What PBM personality assessments can do is assist differential diagnosis of conditions that are conceptualized at least partly in terms of personality characteristics. Some *DSM* diagnostic criteria do refer to state or trait characteristics that PBMs measure well, such as incoherent speech for schizophrenia, depressed mood for MDD, heightened anxiety for PTSD, and impulsivity for ASP. None of these characteristics is specific or unique to these disorders, but their presence increases the likelihood that a person is experiencing them. Although PBM findings thus help to identify the presence and severity of disorders involving maladaptive personality

characteristics, they cannot be expected to show substantial correlations with *DSM* diagnoses, for the reasons just mentioned, nor do PBMs suffice to establish these diagnoses independently of clinical data concerning the nature and duration of a person's problematic symptoms and behaviors.

THE RIM

The RIM consists of 10 standard inkblots printed individually on cards 6 3/4 × 9 1/4 inches (17.15 × 23.50 cm). Five of these blots are printed in shades of gray and black; two of them are in shades of red, gray, and black; and the other three are in shades of various pastel colors. The RIM is administered by showing the cards one at a time and asking, "What might this be?" In a second phase of the examination, the cards are shown again and respondents are asked where in the blots they saw each of their percepts and what made them look as they did. The basic premise underlying Rorschach assessment is that how people look at the inkblots is a representative sample of how they look at the world and that how people look at the world has implications for the kind of person they are.

Rorschach interpretation consists of integrating the structural, thematic, and behavioral sources of data previously described as characterizing PBMs. As a further example of the interpretive significance of a structural Rorschach variable, people whose percepts closely resemble percepts frequently given by most other people are likely to perceive situations and events accurately and in a reasonably conventional manner. As a thematic example, people who frequently report seeing human figures as helping each other in some shared endeavor are likely to feel secure in interpersonal relationships and anticipate that people will interact in collaborative ways. In their test behavior, people who appear self-assured in relating to the examiner and express confidence in the quality of the responses they are giving are likely to be self-assured and confident individuals in their daily lives as well.

History

The RIM was created by Hermann Rorschach, a Swiss psychiatrist who became interested in whether mental hospital patients with different types of disorders would respond differently from each other and from psychologically healthy individuals when asked to say what a series of inkblots looked like. Rorschach eventually selected 10 blots that seemed particularly effective in reflecting individual differences among the patients and nonpatients he examined. He published these 10 blots in a classic book, *Psychodiagnostics* (Rorschach, 1921/1942), and since that time, these same blots have constituted Rorschach's test when used around the world.

However, the standardization of Rorschach's 10 inkblots as the test stimuli did not preclude a proliferation over the years of many different ways of administering, scoring, and interpreting the RIM. In the United States, five distinct Rorschach systems—identified with the names of Sam Beck, Bruno Klopfer, Marguerite Hertz, Zygmunt Piotrowski, and the team of David Rapaport and Roy Schafer—had become well established by the mid-1900s, and a variety of other approaches had become popular in Europe, South America, and Japan. This diversity of methods made it difficult for practitioners to communicate with one another about the implications of their test results and difficult for researchers to cumulate systematic data concerning the psychometric properties and utility of Rorschach findings. Considerable progress was made in resolving this diversity problem when John Exner developed a Comprehensive System (CS) for Rorschach assessment that combined apparently sound elements of previous systems with expanded scoring and interpretive guidelines based on new conceptual formulations and research findings.

Originally published by Exner in 1974 as an antidote to the Rorschach diversity that preceded it, the CS provides instructions for administration and coding that have been widely adopted in the United States and abroad (Exner, 1974, 2003). This standardization improved communication among practitioners, who were now more able than before to speak the same Rorschach language, and it facilitated collaboration among investigators, who were now more able than before to cumulate their research data. As a significant example in this latter regard, an international collaboration of Rorschach investigators generated cumulative reference data on

4,704 adult nonpatients in 21 samples from 17 different countries. In addition to providing normative reference data, this collaboration has demonstrated the worldwide applicability of Rorschach assessment and enriched the study of cross-cultural similarities and differences in personality characteristics (see Meyer, Erdberg, & Shaffer, 2007). As a related note concerning the cross-cultural applicability of Rorschach assessment, comparison studies of U.S. minority groups have shown no substantial CS differences among them. Meyer (2002), for example, found no association between ethnicity and 188 Rorschach summary scores among European American, African American, Asian American, and Native American individuals matched for age, gender, education, marital status, and inpatient status and concluded that "the available data clearly support the cross-ethnic use of the Comprehensive System" (p. 127).

While retaining the standard procedures for Rorschach administration and coding, Exner modified his recommended interpretive strategies over the years to accommodate newly emerging conceptual formulations and research findings. In this sense, the CS has been an evolving rather than a static Rorschach system. Exner died in 2006, and there has since been considerable discussion among Rorschach scholars and practitioners concerning the future of the CS. Will it continue to evolve in light of new ideas and information, and, if so, in whose hands? Will there come instead a time when a new system should be developed to improve on the CS and replace it? If so, who will produce it, who will anoint it as the heir to the CS, and will it be sufficiently widely accepted to become a standard approach and not the precipitant of a new round of nonproductive diversity?

Psychometric Foundations

With respect to its psychometric foundations, extensive research has demonstrated that the RIM can be reliably scored, has adequate retest reliability, and yields valid results when used for its intended purposes. The Rorschach CS comprises over 150 variables and combinations of variables that are coded on the basis of *what* people report seeing in an inkblot (response content); *where* in the inkblot they see it (response location); *why* it looks that way to them (response determinants); and a variety of response elaboration features, such as whether a response has some morbid qualities ("a dead tree stump") or involves aggressive interaction ("people fighting"). Studies in which pairs of examiners have independently coded hundreds of test protocols containing thousands of responses have yielded a median interclass correlation (ICC) of .92, with 95% of the variables showing a level of agreement that would be classified as "excellent" by customary ICC standards (McGrath et al., 2005; Meyer et al., 2002; Viglione & Taylor, 2003).

Numerous test–retest studies over periods of time ranging from a few days to 3 years have demonstrated the substantial stability of Rorschach variables that are conceptualized as relating to trait characteristics of individuals. Retest correlations for most of these variables exceed .75, and some approach .90. The only Rorschach variables that consistently show low retest correlations are variables that are considered to identify states of situational distress, which is a finding that lends construct validity to conceptualizing these variables as indicators of situational rather than persistent phenomena (Exner, 2003, Chapter 11; Grønnerød, 2003, 2006; Viglione & Meyer, 2008).

The most extensive source of information concerning Rorschach validity is a meta-analytic study in which Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999) assessed the validity of 2,276 Rorschach and 5,007 Minnesota Multiphasic Personality Inventory (MMPI) protocols, as measured by correlations of their variables with external (nontest) variables for which there was some reasonable expectation of associations between the test and nontest variables. The obtained average effect sizes in this meta-analysis were .29 for RIM variables and .30 for MMPI variables, thus demonstrating almost identical validity for the two measures. Hiller et al. concluded from their findings that validity evidence for the RIM and MMPI warrants confidence in using both measures for their intended purposes and "is about as good as can be expected from personality tests" (Hiller et al., 1999, p. 291).

Of further significance in the Hiller et al. (1999) findings were some differences between the RIM and MMPI in the strength of their association with two

types of dependent variables. In predicting behavioral outcomes, such as whether patients remain in or drop out of therapy, Rorschach variables were somewhat more effective (with a mean validity coefficient of .37) than MMPI variables (with a mean validity coefficient of .20). In correlating with psychiatric diagnosis and self-reports, on the other hand, the MMPI showed larger effect sizes than the RIM: .37 versus .18. These differences probably reflect a particular sensitivity of the RIM to persistent behavioral dispositions, consistent with the primarily trait implications of most of its variables, and the self-report nature of the MMPI, which resembles the methodology of other self-report measures and includes many of the same kinds of information items on which psychiatric diagnoses are based.

In addition to impressive predictions from Rorschach variables to psychotherapy outcome (see Meyer & Handler, 1997), research findings have demonstrated substantial effect sizes for Rorschach variables in predicting such diverse phenomena as dependency-related behaviors, weight loss among obese persons being treated with behavior modification, military training performance, and both positive attachment behavior and excessive isolation among persons diagnosed with borderline personality disorder (Bornstein, 1999; Elfhag, Rössner, Lindgren, Anderson, & Carlsson, 2004; Fowler, Brunnschweiler, Swales, & Brock, 2005; Hartmann, Sunde, Kristensen, & Martinussen, 2003).

As for psychiatric diagnosis, the earlier observation concerning the indirect nature of PBM diagnostic assessment should not be overlooked. With specific respect to the RIM, it is not a diagnostic test. It cannot by itself directly identify the presence of a *DSM* disorder, and its utility as an assessment instrument should not be judged by its correlations with diagnostic categories. As a personality assessment instrument, on the other hand, the RIM can be expected to assist differential diagnosis by identifying personality characteristics associated with particular psychological disorders. Numerous validation studies have confirmed the incremental validity of Rorschach findings in contributing diagnostic as well as predictive information beyond what can be learned from self-report and interview methods (Blais, Hilsenroth, Castlebury, Fowler, & Baity, 2001; Brand, Armstrong, Loewenstein, & McNary, 2009; Dao, Prevatt, & Heather, 2008; Hartmann, Sunde, Kristensen, & Martinussen, 2003; Janson & Stattin, 2003; Meyer, 2000; Perry, 2001; Porcelli & Mihura, 2010). On the basis of an extensive review of these and other related research findings, the Society for Personality Assessment (2005) issued the following statement:

> Overall, meta-analytic reviews and individual studies show the Rorschach possesses adequate psychometric properties. The research literature consistently demonstrates that the Rorschach can be scored reliably, has scores that measure important psychological functions, and has scores that provide unique information that cannot be obtained from other relevant instruments or clinical interviews. (p. 220)

Despite this evidence, the soundness and utility of Rorschach assessment has frequently been challenged over the years. Current critics persist in questioning the reliability and validity of the RIM, undeterred by repeated evidence-based refutation of their criticisms. As one noteworthy example, it has been argued on the basis of findings from some small and unrepresentative samples that the RIM overpathologizes by incorrectly identifying a large number of nonpatients as being psychologically disturbed (Wood, Nezworski, Garb, & Lilienfeld, 2001; see also Meyer, 2001; Weiner, 2001). To the contrary, in CS normative reference data on a nationally representative sample of 450 nonpatient adults published by Exner and Erdberg in 2005, five key Rorschach indices of psychopathology showed frequencies ranging from just 9% for an elevated Coping Deficit Index ([CDI] >3) down to 4% for an elevated Depression Index ([DEPI] >5), 4% for a positive Hypervigilance Index (HVI), 3% for a lowered Adjusted D Score (AdjD < -1), and 0% for an elevated Perceptual Thinking Index ([PTI] >3). As these data confirm, it is quite rare and far from ordinary for reasonably well-functioning adults to display Rorschach indications of psychopathology.

The overall psychometric soundness of Rorschach assessment notwithstanding, validation

research remains to be done. On the positive side, numerous CS summary scales, such as the PTI (Hilsenroth, Eudell-Simmons, DeFife, & Charnas, 2007), and several widely studied Rorschach scales not included in the CS, such as the Rorschach Oral Dependency Scale (ROD) and the Mutuality of Autonomy Scale (MOA), have been confirmed as dependable measures of the behavioral characteristics they are intended to predict (Bombel, Mihura, & Meyer, 2009; Bornstein & Masling, 2005). On the other hand, not all of the individual variables that constitute these and other scales have been validated as indicators of specific personality characteristics. Just as item analysis in the development of SRIs can lead to modifying or discarding items that correlate poorly with scale scores, so may future CS research call for greater or reduced interpretive emphasis on some of its currently coded variables.

Areas of Application

How people perform on the RIM provides information about a broad range of their personality characteristics, including the adequacy of their cognitive and emotional adaptive capacities; the psychological states and traits that define what they are like as individuals; and the underlying needs, attitudes, conflicts, and concerns that are likely to influence their behavior. Such information about personality functioning helps identify not only the nature and severity of psychological disorder but also whether a person needs and is likely to benefit from various kinds of treatment and the person's prospects for functioning effectively in certain kinds of situations. On this basis, Rorschach findings serve applied purposes by facilitating decisions that are based in part on personality characteristics. Such decision making commonly characterizes the practice of clinical, forensic, and organizational psychology, which are the three areas in which Rorschach assessment is most frequently applied.

Clinical practice. Surveys over the years indicate that the RIM has been the second most frequently used personality instrument in clinical settings, after the MMPI (Hogan, 2005), and the most frequently used personality test in clinical assessments of adolescents (Archer & Newsom, 2000). With respect to differential diagnosis, the RIM, as noted earlier, is not a direct measure of any psychological disorder, but it does identify numerous cognitive, affective, and interpersonal characteristics that are associated with various pathological conditions. For example, dependable Rorschach indices of thinking disorder (elevated *WSum6*) and poor reality testing (high *X-%*) can be helpful in identifying schizophrenia spectrum disorders; indices of dysphoria (elevated *C′*, *Col-Shd Blds*) can help calibrate the depth of a depressive condition; and indices of hypervigilance and interpersonal aversion (elevated HVI) can suggest the presence of a paranoid frame of reference. These and other applications of Rorschach findings in differential diagnosis are elaborated in an extensive literature (see Exner & Erdberg, 2005; Hartmann, Norbech, & Grønnerød, 2006; Huprich, 2006; Kleiger, 1999; Weiner, 2003b).

Two current developments in approaches to classification may expand the role of Rorschach assessment in differential diagnosis. The recently formulated *Psychodynamic Diagnostic Manual* (PDM; PDM Task Force, 2006) places greater emphasis on personality patterns (P axis) and on strengths and limitations in mental functioning (M axis) than on the symptom patterns (S axis) that dominate *DSM* diagnostic criteria. If the PDM fosters diagnostic practices that are more closely tied to personality characteristics than the *DSM*, then personality assessment instruments are likely to play an increased role in arriving at formal diagnoses. As a second development, this same role expectation applies to the extent to which the fifth edition of the *DSM* currently in preparation depends more heavily than the *DSM–IV–TR* on personality characteristics as criteria for diagnostic categorization.

Rorschach assessment contributes to clinical practice not only by assisting in differential diagnosis of psychopathology but also by facilitating treatment planning and outcome evaluation. With respect to treatment planning, Rorschach findings can inform key decisions before and during a therapeutic intervention. In the decision of whether to recommend psychotherapy, for example, some predictive validity derives from the fact that certain personality characteristics measured by Rorschach variables are typically associated with the ability to

participate in and benefit from psychological treatment. These include being open to experience (*Lambda* not elevated), cognitively flexible (balanced *a:p* ratio), emotionally responsive (adequate *WSumC* and *Afr*), psychologically minded ($FD > 0$), and interpersonally receptive ($T > 0$, $SumH > 2$), each of which facilitates engagement and progress in therapy (see Clarkin & Levy, 2004; Fowler et al., 2004; Weiner & Bornstein, 2009, Chapter 2).

With respect to treatment selection, the personality style and severity of distress or disorganization revealed by Rorschach findings can help indicate whether a person's treatment needs will best be met by a supportive approach oriented toward relieving distress, a cognitive–behavioral approach designed to modify symptoms or behavior, or an exploratory approach intended to enhance self-understanding. Rorschach assessment has also proved useful in monitoring progress in therapy and evaluating its effectiveness, particularly when information about personality strengths and weaknesses that is obtained during the course of treatment or following termination can be compared with baseline testing information (see Bram, 2010; Rothschild, Lacoua, Eshel, & Stein, 2008; Weiner, 2005b; Weiner & Exner, 1991).

Finally, of note, the essence of the RIM as a personality assessment instrument limits what it can be expected to do in clinical applications as well as in the forensic and organizational applications discussed next, particularly with respect to identifying past events or predicting future ones. Only when there is a substantial correlation between specific personality characteristics and the likelihood of people having had certain experiences or behaved in certain ways can Rorschach data provide dependable indications of whether such experiences or behaviors have occurred. Thus, for example, Rorschach assessment cannot identify whether a child has been sexually abused or an adult has had a substance abuse problem. The predictive validity of Rorschach findings is similarly constrained by the extent to which personality characteristics determine whatever behavior is being predicted. To put this into consideration the other way around, the more a future behavior is a function of situational factors and individual characteristics unrelated to dimensions of personality, the less helpful the RIM and other personality assessment instruments will be in predicting it (see Weiner, 2003a).

Forensic practice. Forensic Rorschach applications derive from the implications of certain personality characteristics for judicial decision making in criminal, civil, and family law cases. In criminal cases, for example, impaired reality testing and inability to think logically and coherently (as indicated by an elevated PTI) often has a bearing on whether accused persons should be considered competent to proceed to trial (which, in legal terms, consists in part of having a factual and rational understanding of the charges against them) or should be held responsible for their alleged criminal behavior (which, in legal terms, consists in part of being able to appreciate the wrongfulness of their conduct). In civil law cases involving personal injury suits for damages, the elevation of Rorschach indices of anxiety (e.g., minus *D* score), depression (e.g., Depression Index), and disturbing thoughts, especially about being vulnerable to bodily harm (e.g., morbid contents) can help determine the extent to which a person has become emotionally distressed as a consequence of alleged irresponsible behavior on the part of some person or entity.

In family law cases, courts make their determinations of how a child's time and supervision should be divided between separated or divorced parents partly on the basis of information about the personality characteristics of the child and the child's parents. Although there are no infallible guidelines concerning which of two persons would be the better parent for a particular child, certain personality characteristics that are measured by the RIM are likely to enhance or detract from parents' ability to meet the needs of their children. These characteristics include, but are not limited to, whether a parent gives little or considerable evidence of serious psychological disturbance; whether a parent's coping skills and stress tolerance seem ample or limited; and whether the parent appears to be an interpersonally accessible and receptive person or a withdrawn and uncommunicative person. Publications by Erard (2005), Gacono and Evans (2008), and Weiner (2006, 2007), among others, elaborate these

and other substantive guidelines for forensic Rorschach assessment in criminal, personal injury, and child custody cases.

The established place of Rorschach assessment in forensic settings is reflected in survey findings that over one third (36%) of forensic psychologists use the RIM in their practice (Archer, Buffington-Vollum, Stredny, & Handel, 2006). Nevertheless, some critics of Rorschach assessment have asserted that the RIM should not be used in forensic cases and that Rorschach testimony will not be accepted into evidence in the courtroom (Grove, Barden, Garb, & Lilienfeld, 2002). To the contrary, actual court records indicate that Rorschach testimony is almost always welcome in the courtroom. Over a 50-year period between 1945 and 1995, Rorschach testimony was presented to a U.S. federal, state, or military court of appeals in a total of 247 cases, and a record review of these cases indicated that this testimony was accepted into evidence without challenge in 90% of them (Meloy, Hansen, & Weiner, 1997). In the years between 1996 and 2006, a total of 150 appellate cases involved Rorschach testimony, and only three (2%) of these cases included recorded criticism of this testimony (Meloy, 2008). The 1996–2006 tripling of the 1945–1995 annual rate of appellate Rorschach testimony (150 over a 10-year span vs. 247 during a 50-year period) and the much reduced frequency of challenges (from 10% to 2%) speak to the contemporary growth and acceptability of forensic Rorschach assessment.

Organizational practice. In organizational practice, Rorschach assessment is utilized primarily to assist in the selection and evaluation of personnel. Decisions about the suitability of individuals for a position in an organization or for promotion within the organization to a leadership position usually involve (a) identifying some personality requirements for success in the position and (b) determining the extent to which a candidate shows these personality characteristics. For example, among persons being considered for hire as an air traffic controller or nuclear power plant supervisor, their candidacy would be supported by indications on personality testing of good coping capacities and the ability to remain calm and exercise good judgment even in highly stressful situations—the dependable Rorschach indications of which would be a high *EA*, $D \geq 0$, and *XA%* in the normal range. In a detailed presentation of Rorschach scales relevant to occupational performance, Del Giudice (2010) concluded that "the Rorschach may represent a unique and potentially valuable tool for assessing personality as part of comprehensive personality selection procedures" (p. 78).

The TAT

The TAT is a storytelling technique in which respondents are asked to make up stories about pictures of people and scenes. Respondents are instructed that their stories should have a beginning, a middle, and an end and should include what is happening in the picture, what led up to this situation, what the people in the picture are thinking and feeling, and what the outcome of the situation will be. When they have finished telling their story about a picture, respondents are asked to add any story elements they have omitted (e.g., "How is the man in the picture feeling?" "What is going to happen next?"). The full set of TAT pictures comprises 31 achromatic cards measuring 9 1/4 × 11 inches (23.50 × 27.94 cm), some of which are intended for use with boys or girls ages 4 to 14 and some of which are intended for use with male and female respondents aged 15 or older. In typical practice, examiners administer nine to 12 of the cards, which are selected on the basis of their seeming relevance to salient issues in the life of the person being assessed.

The structural data in TAT protocols consist of such objective story features as their length, the amount of detail they contain, and the picture elements they mention. Story length can provide information about whether people are approaching this task, and perhaps other situations in their lives as well, in a relatively open and revealing fashion (long stories) or in a relatively guarded manner that conceals more than it reveals (short stories). The detail in TAT stories can vary from a precisely specified account of who is doing what to whom and why (which might reflect some obsessive–compulsive characteristics) to a vague and superficial description of people (which might suggest a shallow style of dealing with affective and interpersonal

experience). Stories that make no mention of prominent and frequently noticed elements of the pictures may identify tendencies to be insufficiently attentive to what is obvious and important in situations, and stories that focus on rarely mentioned peripheral details may identify tendencies to become preoccupied with what is obscure and of little significance.

Like the thematic imagery that emerges in Rorschach responses, the content of TAT stories often provides clues to a person's inner life. Because many of the TAT cards depict people, they give respondents numerous opportunities to describe the aspirations and intentions, attributes and shortcomings, hopes and fears, past and future life experiences, and other characteristics of the people in their stories; such descriptions are likely to reveal aspects of how storytellers view themselves, other people, life events, and what the future holds for them. Also in common with other performance-based measures of personality, TAT administration provides behavioral data suggesting how respondents are typically likely to approach task-oriented and interpersonal situations. For example, those who present their stories in an engaged and self-assured manner and relate to the examiner in an assertive but friendly fashion are likely to be similarly engaged, self-assured, assertive, and friendly on other occasions as well. Those who respond in a detached and tentative fashion and strike the examiner as surly or deferential are likely to be providing a glimpse of a general disposition to deal with tasks and people in these ways.

History

The TAT was devised by Henry Murray, a physician, biochemist, and psychoanalyst who was director of the Harvard Psychological Clinic from 1928 to 1943. Two threads in Murray's professional life influenced his development and promulgation of TAT assessment. First, during the 1930s, he became intrigued with the notion that stories people tell, particularly when these stories are products of their imagination, can reveal many of their underlying thoughts and feelings. In collaboration with an artistic colleague, Christiana Morgan, he began searching for pictures that he thought could be used to good effect as a stimulus for eliciting stories rich in personal meaning (see Morgan & Murray, 1935).

Second, Murray became convinced that personality research should focus on each individual's unique integration of psychological characteristics, rather than on the general nature of these characteristics, and should explore the individual experience and kinds of lives that people lead instead of the nature of particular personality characteristics. To this end, Murray's first major research project was an intensive psychological study of 50 male Harvard students, each of whom was assessed individually with over 20 different procedures, including his newly developed picture—story method. The results of this study were published in 1938, *Explorations in Personality: A Clinical and Experimental Study of Fifty Men of College Age*, a classic book that is best known for Murray's presentation of his idiographic approach to studying people and personality functioning but in which he also elaborated for the first time how the TAT could be applied in conjunction with other assessment methods to gain insight into the influences that shape an individual's personality. After some modifications of the picture set used in the Harvard Clinic study, the final 31-card version of the TAT, which remains the version in use today, was published in 1943 (Murray, 1943/1971).

Murray developed an elaborate system for coding TAT stories, but neither his system nor any of numerous other global scoring schemes proposed over the years became widely accepted by either researchers or practitioners (see Jenkins, 2008). Instead, TAT interpretation has most often followed guidelines for an "inspection technique" recommended by Leopold Bellak (Bellak, 1947; Bellak & Abrams, 1997). The inspection technique begins with examining an individual's stories for repetitive themes and recurring elements that appear to fall together in meaningful ways. Interpretation then proceeds by reading through a person's stories with close attention to how the people in the stories are described and interact, how the story plots begin and end, the emotional tone that characterizes the stories, and the cognitive integrity of the narrative.

As an illustration of the implications of the characteristics that respondents attribute to people and events in their TAT stories, a story about a boy who is musically talented, practices the violin diligently, wants to become a famous performer, and

eventually achieves this objective would suggest attitudes of confidence in one's capabilities, commitment to working hard toward ambitious goals, and expectations of success in what one tries to accomplish. By contrast, a story about a "feeble" man who is attempting to climb a rope "but isn't strong enough to pull himself up" would suggest self-perception as a weak or ineffectual person, expectations of little to be gained from trying hard, and a future fraught with failure. Whether people interacting in TAT stories are being helpful or hurtful, giving praise or criticism, or reaching out to or rebuffing one another may say something about the storyteller's fears or expectations in interpersonal situations.

The emotional tone of the TAT stories that people tell usually has some implications for their affective state or disposition. Thus, recurrent descriptions of story characters as being happy or sad, anxious or relaxed, regretful or self-satisfied, or angry or at peace with the world suggest that the person may be experiencing or be prone to similar affects. Additionally, the circumstances in the stories that appear to have aroused these affects may identify the kinds of situations that cause the person to feel this way, and how story characters deal with these circumstances often reflect a person's preferred coping style (e.g., whether with passive withdrawal from problematic situations or active efforts to overcome or resolve them).

As for the cognitive integrity of TAT narratives, incoherent stories that are difficult to follow may reflect loose or dissociated thinking, and stories with highly unlikely outcomes can signify poor judgment about how one event leads to another and, thus, be an indication of faulty reality testing. Sources of additional information on TAT interpretation include contributions by Bellak and Abrams (1997), Cramer (1996), Henry (1956), Teglasi (2001), and Weiner and Greene (2008, Chapter 12).

Finally, of note in the development of the TAT are three well-conceived content rating scales that have been validated as measures of specific dimensions of personality functioning. The Social Cognition and Object Relations Scale (SCORS) was designed by Westen and colleagues (Westen, 1991; Westen, Lohr, Silk, Gold, & Kerber, 1990) to tap underlying attitudes that people have toward themselves, toward other people, and toward social relationships. The SCORS system rates stories on eight dimensions of object relatedness for the level of maturity reflected in the actions and attitudes of the depicted characters. The SCORS has been validated for a variety of applications, particularly in clinical practice (e.g., Kelly, 2007; Peters, Hilsenroth, Eudell-Simmons, Blagys, & Handler, 2006; Porcerelli et al., 2006).

The Defense Mechanism Manual (DMM), developed by Cramer (1991), identifies seven story characteristics considered to reflect the immature and maladaptive defense mechanism of *denial* (failing to see or ignoring something that is really there); the less immature but still usually maladaptive mechanism of *projection* (attributing characteristics to people and situations in the absence of adequate justification); and the relatively mature and adaptive mechanism of *identification* (adopting certain characteristics of other people in an attempt to become like them). The frequency with which a person's TAT stories reflect these defenses provides an index of the individual's overall level of defensiveness, and the relative magnitude of the subtotal scores for denial, projection, and identification indicates the maturity and likely adaptive value of his or her defense preferences (Cramer, 2009; Hibbard, Porcerelli, Kamoo, Schwartz, & Abell, 2010).

Whereas the SCORS and DMM have both clinical and research applications, the third notable TAT scale, the Need for Achievement (n-Ach) scale developed by McClelland and his colleagues (McClelland, Atkinson, Clark, & Lowell, 1953; McClelland, Clark, Roby, & Atkinson, 1958), has served primarily as a research tool. The n-Ach scale codes stories for six presumably achievement-related features, such as a depicted character desiring to reach some goal, engaging in activity intended to reach that goal, and anticipating success or failure in reaching the goal. Positive demonstrations of the utility of the n-Ach scale as a conceptually rich and empirically sturdy research tool fostered numerous experimental applications of the TAT and the construction of similar companion scales for measuring needs for affiliation, power, intimacy, and responsibility. This experimental work and the nature of these companion scales are reviewed by Cramer (1996, Chapter 15) and McClelland (1999).

Psychometric Foundations

Aside from a widely used and fairly standard set of instructions based on Murray's original guidelines, research and practice with the TAT has been largely unsystematic, particularly with respect to which cards and how many of the cards are shown and in what order. This variability, together with the primarily qualitative approach that typifies TAT interpretation in clinical practice, has made it difficult to generate the kinds of quantitative data that facilitate determining the reliability of an assessment instrument and the validity of its scores for various purposes. The TAT's lack of systematization and its limited psychometric verification along traditional lines have fueled a long history of controversy between assessment specialists who value the utility of the instrument and critics who have questioned the propriety of using the TAT in clinical practice (see Cramer, 1999; Hibbard, 2003; Lilienfeld, Wood, & Garb, 2000).

With respect to this debate, three important considerations bear on showing how and why the TAT can be used effectively for certain purposes. First, evidence supporting the validity of test scores typically has been questioned on the basis of low correlations with clinical diagnoses and SRI data. Like the RIM, however, the TAT is a personality instrument, not a diagnostic test. Hence, it may provide diagnostically useful information about conditions that are defined at least in part by personality characteristics. However, the TAT does not directly identify any specific disorders, nor should its validity evidence be judged by its correlations with specific disorders. As for comparing TAT findings with SRI data, this chapter began by identifying SRIs and PBMs as types of tests that differ in how they are constructed, the kinds of responses they require, the amount of structure they provide, and the levels of conscious awareness they tap.

Because of these differences between performance-based and self-report methods, the validity of the former cannot properly be determined from its correlations with the latter. As a more general dictum in these regard, the key data for validating any assessment instrument come not from their correlations with other assessment instruments, all of which are inferential measures, but from objective facts and observable behavior. Compelling evidence of criterion validity emerges when personality test scores can be shown to correlate with external (non-test) variables consisting of what people are in fact like and how they are observed to behave.

Second, the predominantly qualitative approach to TAT interpretation in clinical practice has been supplemented with quantitative scales that are accessible to psychometric verification. Research with the previously mentioned SCORS, DMM, and n-Ach scale has demonstrated that TAT assessment can be objectified to yield reliable and valid scales for measuring dimensions of personality functioning. Additionally, almost 20 other TAT scales reviewed in the previously cited text by Jenkins (2008) have shown respectable psychometric properties in at least a few research studies, although none has become widely used.

Third, the primary purpose of TAT assessment is to generate hypotheses about a person's inner life and coping style. The value of the TAT resides in the extent to which these hypotheses help examiners understand the people they evaluate and arrive at useful conclusions and recommendations concerning them. Even when not definitive in themselves, TAT-generated hypotheses can provide valuable clues to possibilities that should be explored further. Psychologists concerned that this qualitative perspective detracts from the scientific status of assessment psychology should keep in mind that generating hypotheses is as much a part of science as confirming hypotheses.

Areas of Application

Because the TAT functions best as a measure of underlying needs, attitudes, conflicts, and concerns, its primary application is in clinical work, mostly in planning, conducting, and monitoring progress in psychotherapy. Over the years, the TAT has been the third most widely used personality assessment measure in clinical settings, after the MMPI and the RIM (Hogan, 2005). TAT findings can be especially helpful in evaluating people who are seeking mental health care but are unable or disinclined to reveal very much about themselves. The test data in such instances typically go beyond interview data in illuminating issues that should be addressed in

psychotherapy, including (a) the types of concerns that need to be resolved for the person to feel better and function more effectively, (b) the kinds of attitudes the person has toward key figures in his or her life and toward interpersonal relatedness, and (c) the sorts of situations that are likely to be distressing or gratifying to the person and how the person responds to these situations.

With respect to planning and conducting psychotherapy, TAT-generated hypotheses about a person's inner life can help therapists identify treatment goals, decide on the timing and focus of their interventions, and anticipate obstacles to progress in establishing a productive working alliance and effecting positive behavior change. As for monitoring progress in psychotherapy, both the SCORS and the DMM have demonstrated utility as indices of improvement during the course of psychological treatment (Cramer & Blatt, 1990; Fowler et al., 2004; Thompson-Brenner, Boisseau, & Satir, 2010).

Because *DSM* diagnosis distinguishes among disorders primarily on the basis of manifest symptoms and past and current behavior, as noted previously, and gives little attention to a person's underlying attitudes and concerns, the richness of the TAT in generating hypotheses about a person's inner life seldom contributes as much to differential diagnostic evaluations as it does to treatment planning and evaluation. Nevertheless, there are occasions in which certain structural, thematic, and behavioral features of a TAT protocol can reinforce and firm up diagnostic impressions based on other sources of information. Structurally, for example, disjointed narratives may reflect disordered thinking; thematically, repetitive themes of suspicion and betrayal may suggest paranoia; behaviorally, lengthy stories delivered more rapidly than the examiner can record them may indicate such hypomanic features as racing thoughts.

Although stories told to TAT pictures are better suited for generating hypotheses than for establishing the reasonable certainty expected in the courtroom, thematic preoccupations may have forensic applications should they appear to document a state of mind relevant to a legal issue. In a personal injury case, for example, TAT stories reflecting pervasive fears of being harmed or damaged might lend weight to a claim of PTSD that is central to a litigant's quest for damages. In organizational practice, some TAT indices of specific personality characteristics may prove valuable in making decisions about the selection or promotion of personnel. For example, meta-analytic studies have shown a statistically significant average effect size for the n-Ach scale in predicting such outcomes as income earned, job performance, and successful entrepreneurship (Collins, Hanges, & Locke, 2004; Spangler, 1992).

SUMMARY

PBMs of personality are indirect assessments of an individual's psychological states, traits, and underlying attitudes and concerns. Unlike SRIs, which sample fairly directly what people are able and willing to say about themselves, PBMs infer characteristics of people from how they perform on various tasks. Inferences based on PBMs are often more speculative than conclusions based on SRI responses, but PBMs are more likely than SRIs to reveal characteristics of which people are unaware, and they are less susceptible than SRIs to impression management. PBMs provide information about a broad range of personality characteristics, but they do not assess psychopathology directly. Instead, they help to identify the presence and severity of disorders that are conceptualized as involving distinctive personality characteristics. The most widely used personality PBMs, each with its particular format, history, psychometric foundations, and areas of application, are the RIM, in which respondents are asked to say what a series of inkblots might be, and the TAT, in which respondents are asked to make up stories about a series of pictures.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94. doi:10.1207/s15327752jpa8701_07

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update.

Assessment, 7, 227–235. doi:10.1177/1073191 10000700303

Bellak, L. (1947). *A guide to the interpretation of the Thematic Apperception Test*. New York, NY: Psychological Corporation.

Bellak, L., & Abrams, D. M. (1997). *The TAT, CAT, and SAT in clinical use* (3rd ed.). Boston, MA: Allyn & Bacon.

Beutler, L. E., & Groth-Marnat, G. (2003). *Integrative assessment of adult personality* (2nd ed.). New York, NY: Guilford Press.

Blais, M. A., Hilsenroth, M. J., Castlebury, F., Fowler, J. C., & Baity, M. R. (2001). Predicting DSM-IV Cluster B personality disorder criteria from MMPI-2 and Rorschach data. *Journal of Personality Assessment, 76*, 150–168. doi:10.1207/S15327752JPA7601_9

Bombel, G., Mihura, J. L., & Meyer, G. F. (2009). An examination of the construct validity of the Rorschach Mutuality of Autonomy (MOA) Scale. *Journal of Personality Assessment, 91*, 227–237. doi:10.1080/00223890902794267

Bornstein, R. F. (1999). Criterion validity of objective and projective dependency tests: A meta-analytic assessment of behavioral prediction. *Psychological Assessment, 11*, 48–57. doi:10.1037/1040-3590. 11.1.48

Bornstein, R. F., & Masling, J. M. (Eds.). (2005). The Rorschach oral dependency scale. In *Scoring the Rorschach: Seven validated systems* (pp. 135–158). Mahwah, NJ: Erlbaum.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63*, 363–386. doi:10.1207/ s15327752jpa6302_14

Bram, A. D. (2010). The relevance of the Rorschach and patient-examiner relationships in treatment planning and outcome assessment. *Journal of Personality Assessment, 92*, 91–115. doi:10.1080/00223890903508112

Brand, B. L., Armstrong, J. G., Loewenstein, R. J., & McNary, S. W. (2009). Personality differences on the Rorschach of dissociative identity disorder, borderline personality disorder, and psychotic inpatients. *Psychological Trauma: Theory, Research, Practice, and Policy, 1*, 188–205. doi:10.1037/a0016561

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Clarkin, J. F., & Levy, K. N. (2004). The influence of client variables on psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 194–236). Hoboken, NJ: Wiley.

Collins, C. J., Hanges, P. J., & Locke, E. A. (2004). The relationship of achievement motivation to entrepreneurial behavior: A meta-analysis. *Human Performance, 17*, 95–117. doi:10.1207/S15327043HUP1701_5

Cramer, P. (1991). *The development of defense mechanisms: Theory, research, and assessment*. New York, NY: Springer-Verlag. doi:10.1007/978-1-4613-9025-1

Cramer, P. (1996). *Storytelling, narrative, and the Thematic Apperception Test*. New York, NY: Guilford Press.

Cramer, P. (1999). Future directions for the Thematic Apperception Test. *Journal of Personality Assessment, 72*, 74–92. doi:10.1207/s15327752jpa7201_5

Cramer, P. (2009). The development of defense mechanisms from pre-adolescence to early adulthood: Do IQ and social class matter? A longitudinal study. *Journal of Research in Personality, 43*, 464–471. doi:10.1016/j.jrp.2009.01.021

Cramer, P., & Blatt, S. J. (1990). Use of the TAT to measure change in defense mechanisms following intensive psychotherapy. *Journal of Personality Assessment, 54*, 236–251. doi:10.1207/s15327752jpa5401&2_23

Dana, R. H. (2005). *Multicultural assessment: Principles, applications, and examples*. Mahwah, NJ: Erlbaum.

Dao, T. K., Prevatt, F. H., & Heather, L. (2008). Differentiating psychotic from nonpsychotic patients with the MMPI-2 and Rorschach. *Journal of Personality Assessment, 90*, 93–101. doi:10.1080/00223890 701693819

Del Giudice, M. J. (2010). What might this be? Rediscovering the Rorschach as a toll for personnel selection in organizations. *Journal of Personality Assessment, 92*, 78–89. doi:10.1080/00223890 903382385

Elfhag, K., Rössner, S., Lindgren, T., Andersson, I., & Carlsson, A. M. (2004). Rorschach personality predictors of weight loss with behavior modification on obesity treatment. *Journal of Personality Assessment, 83*, 293–305. doi:10.1207/s15327752jpa8303_11

Erard, R. E. (2005). What the Rorschach can contribute to child custody and parenting time evaluations. *Journal of Child Custody*, *2*, 119–142. doi:10.1300/ J190v02n01_07

Exner, J. E., Jr. (1974). *The Rorschach: A comprehensive system*. New York, NY: Wiley.

Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system*: *Vol. 1. Basic foundations and principles of interpretation* (4th ed.). Hoboken, NJ: Wiley.

Exner, J. E., Jr., & Erdberg, P. (2005). *The Rorschach: A comprehensive system*: *Vol. 2. Advanced interpretation* (3rd ed.). Hoboken, NJ: Wiley.

Fowler, J. C., Ackerman, S. J., Speanburg, S., Bailey, A., Blagys, M., & Conklin, A. C. (2004). Personality

and symptom change in treatment-refractory inpatients: Evaluation of the phase model of change using Rorschach, TAT, and DSM-IV Axis V. *Journal of Personality Assessment, 83*, 306–322. doi:10.1207/s15327752jpa8303_12

Fowler, J. C., Brunnschweiler, B., Swales, S., & Brock, J. (2005). Assessment of Rorschach dependency measures in female inpatients diagnosed with borderline disorder. *Journal of Personality Assessment, 85*, 146–153. doi:10.1207/s15327752jpa8502_07

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology: Interdisciplinary and Applied, 8*, 389–413. doi:10.1080/00223980.1939.9917671

Gacono, C. B., & Evans, F. B. (Eds.). (2008). *Handbook of forensic Rorschach assessment.* New York, NY: Routledge.

Grønnerød, C. (2003). Temporal stability of the Rorschach method: A meta-analytic review. *Journal of Personality Assessment, 80*, 272–293. doi:10.1207/S15327752JPA8003_06

Grønnerød, C. (2006). Reanalysis of the Grønnerød (2003) Rorschach temporal stability meta-analysis set. *Journal of Personality Assessment, 86*, 222–225. doi:10.1207/s15327752jpa8602_12

Grove, W. M., Barden, R. C., Garb, H. N., & Lilienfeld, S. O. (2002). Failure of Rorschach-Comprehensive-System-based testimony to be admissible under the Daubert-Joiner-Kumho standards. *Psychology, Public Policy, and Law, 8*, 216–234. doi:10.1037/1076-8971.8.2.216

Hartmann, E., Norbech, P. B., & Grønnerød, C. (2006). Psychopathic and nonpsychopathic violent offenders on the Rorschach: Discriminative features and comparisons with schizophrenic inpatient and university student samples. *Journal of Personality Assessment, 86*, 291–305. doi:10.1207/s15327752jpa8603_05

Hartmann, E., Sunde, T., Kristensen, W., & Martinussen, M. (2003). Psychological measures as predictors of military training performance. *Journal of Personality Assessment, 80*, 87–98. doi:10.1207/S15327752JPA8001_17

Henry, W. E. (1956). *The analysis of fantasy: The thematic apperception technique in the study of personality.* New York, NY: Wiley.

Hibbard, S. (2003). A critique of Lilienfeld et al.'s (2000) "The Scientific Status of Projective Techniques." *Journal of Personality Assessment*, *80*, 260–271. doi:10.1207/S15327752JPA8003_05

Hibbard, S., Porcerelli, J., Kamoo, R., Schwartz, M., & Abell, S. (2010). Defense and object relational maturity on Thematic Apperception Test scales indicate levels of personality organization. *Journal of Personality Assessment, 92*, 241–253. doi:10.1080/00223891003670190

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach validity. *Psychological Assessment, 11*, 278–296. doi:10.1037/1040-3590.11.3.278

Hilsenroth, M. J., Eudell-Simmons, E. R., DeFife, J. A., & Charnas, J. W. (2007). The Rorschach Perceptual Thinking Index (PTI): An examination of reliability, validity, and diagnostic efficiency. *International Journal of Testing, 7*, 269–291. doi:10.1080/15305050701438033

Hogan, T. P. (2005). 50 widely used psychological tests. In G. P. Koocher, J. C. Norcross, & S. S. Hill, III (Eds.), *Psychologists' desk reference* (2nd ed., pp. 101–104). New York, NY: Oxford University Press.

Huprich, S. K. (Ed.). (2006). *Rorschach assessment of the personality disorders.* Mahwah, NJ: Erlbaum.

Janson, H., & Stattin, H. (2003). Predictions of adolescent and adult delinquency from childhood Rorschach ratings. *Journal of Personality Assessment, 81*, 51–63. doi:10.1207/S15327752JPA8101_05

Jenkins, S. R. (2008). *Handbook of clinical scoring systems for thematic apperceptive techniques.* New York, NY: Erlbaum.

Kelly, F. D. (2007). The clinical application of the Social Cognition and Object Relations Scale with children and adolescents. In S. R. Smith & L. Handler (Eds.), *The clinical assessment of children and adolescents* (pp. 169–184). Mahwah, NJ: Erlbaum.

Kleiger, J. H. (1999). *Disordered thinking and the Rorschach.* Hillsdale, NJ: Analytic Press.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.

McClelland, D. C. (1999). How the test lives on: Extensions of the Thematic Apperception Test approach. In L. Gieser & M. I. Stein (Eds.), *Evocative images: The Thematic Apperception Test and the art of projection* (pp. 163–175). Washington, DC: American Psychological Association. doi:10.1037/10334-012

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive.* New York, NY: Appleton-Century-Crofts. doi:10.1037/11144-000

McClelland, D. C., Clark, R. A., Roby, T. B., & Atkinson, J. W. (1958). The effect of the need for achievement on thematic apperception. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society* (pp. 64–82). Princeton, NJ: Van Nostrand.

McCrae, R. R., & Terracciano, A. (2008). The five-factor model and its correlates in individuals and cultures. In F. J. R. van de Vijer, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 249–283). New York, NY: Taylor & Francis.

McGrath, R. E., Pogge, D. L., Stokes, J. M., Cragnolino, A., Zaccario, M., Hayman, J., . . . Wayland-Smith, D. (2005). Field reliability of Comprehensive System scoring in an adolescent inpatient sample. *Assessment*, *12*, 199–209. doi:10.1177/1073191104273384

Meloy, J. R. (2008). The authority of the Rorschach: An update. In C. B. Gacono & F. B. Evans (Eds.), *Handbook of forensic Rorschach assessment* (pp. 79–87). New York, NY: Routledge.

Meloy, J. R., Hansen, T., & Weiner, I. B. (1997). Authority of the Rorschach: Legal citations in the past 50 years. *Journal of Personality Assessment, 69*, 53–62. doi:10.1207/s15327752jpa6901_3

Meyer, G. J. (2000). The incremental validity of the Rorschach Prognostic Rating Scale over the MMPI Ego Strength Scale and IQ. *Journal of Personality Assessment, 74*, 356–370. doi:10.1207/S15327752JPA7403_2

Meyer, G. J. (2001). Evidence to correct misperceptions about Rorschach norms. *Clinical Psychology: Science and Practice, 8*, 389–396. doi:10.1093/clipsy.8.3.389

Meyer, G. J. (2002). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment, 78*, 104–129. doi:10.1207/S15327752JPA7801_07

Meyer, G. J., Erdberg, P., & Shaffer, T. W. (2007). Toward international normative reference data for the Comprehensive System. *Journal of Personality Assessment, 89*(Suppl.), 201–216.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: Meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 69*, 1–38. doi:10.1207/s15327752jpa6901_1

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fow1er, J. C., Piers, C. C., & Resnick, J. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78*, 219–274. doi:10.1207/S15327752JPA7802_03

Meyer, G. J., & Kurtz, J. E. (2006). Advancing personality assessment terminology: Time to retire "objective" and "projective" as personality test descriptors. *Journal of Personality Assessment, 87*, 223–225. doi:10.1207/s15327752jpa8703_01

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry, 34*, 289–306.

Murray, H. A. (1938). *Explorations in personality: A clinical and experimental study of fifty men of college age.* New York, NY: Oxford University Press.

Murray, H. A. (1971). *Thematic Apperception Test: Manual.* Cambridge, MA: Harvard University Press. (Original work published 1943)

Perry, W. (2001). Incremental validity of the Ego Impairment Index: A reexamination of Dawes (1999). *Psychological Assessment, 13*, 403–407. doi:10.1037/1040-3590.13.3.403

Peters, E. J., Hilsenroth, M. J., Eudell-Simmons, E. M., Blagys, M. D., & Handler, L. (2006). Reliability and validity of the Social Cognition and Object Relations Scale in clinical use. *Psychotherapy Research, 16*, 617–626. doi:10.1080/10503300600591288

Porcelli, P., & Mihura, J. L. (2010). Assessment of alexithymia with the Rorschach Comprehensive System: The Rorschach Alexithymia Scale (RAS). *Journal of Personality Assessment*, *92*, 128–136. doi:10.1080/00223890903508146

Porcerelli, J. H., Shahar, G., Blatt, S. J., Ford, R. Q., Mezza, J. A., & Greenlee, L. M. (2006). Social Cognition and Object Relations Scale: Convergent validity and changes following intensive inpatient treatment. *Personality and Individual Differences, 41*, 407–417. doi:10.1016/j.paid.2005.10.027

Psychodynamic Diagnostic Manual Task Force. (2006). *Psychodynamic diagnostic manual.* Silver Spring, MD: Alliance of Psychoanalytic Organizations.

Rorschach, H. (1942). *Psychodiagnostics: A diagnostic test based on perception.* Bern, Switzerland: Hans Huber. (Original work published 1921)

Rothschild, L., Lacoua, L., Eshel, Y., & Stein, D. (2008). Changes in defensiveness and in affective distress following inpatient treatment of eating disorders: Rorschach Comprehensive System and self-report measures. *Journal of Personality Assessment, 90*, 356–367. doi:10.1080/00223890802107982

Society for Personality Assessment. (2005). The status of the Rorschach in clinical and forensic practice: An official statement by the Board of Trustees of the Society for Personality Assessment. *Journal of Personality Assessment, 85*, 219–237. doi:10.1207/s15327752jpa8502_16

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin*, *112*, 140–154. doi:10.1037/0033-2909.112.1.140

Teglasi, H. (2001). *Essentials of TAT and other storytelling techniques assessment.* New York, NY: Wiley.

Thompson-Brenner, H., Boisseau, C. L., & Satir, D. A. (2010). Adolescent eating disorders: Treatment and response in a naturalistic study. *Journal of Clinical Psychology, 66*, 277–301.

Viglione, D. J., & Meyer, G. J. (2008). An overview of Rorschach psychometrics for forensic practice. In C. B. Gacono & F. B. Evans (Eds.), *Handbook of forensic Rorschach assessment* (pp. 21–53). New York, NY: Routledge.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach comprehensive system coding. *Journal of Clinical Psychology, 59*, 111–121. doi:10.1002/jclp.10121

Weiner, I. B. (2001). Considerations in collecting Rorschach reference data. *Journal of Personality Assessment, 77*, 122–127. doi:10.1207/S15327752JPA7701_08

Weiner, I. B. (2003a). Prediction and postdiction in clinical decision making. *Clinical Psychology: Science and Practice, 10*, 335–338. doi:10.1093/clipsy.bpg030

Weiner, I. B. (2003b). *Principles of Rorschach interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.

Weiner, I. B. (2005a). Integrative personality assessment with self-report and performance-based measures. In S. Strack (Ed.), *Handbook of personology and psychopathology* (pp. 317–331). Hoboken, NJ: Wiley.

Weiner, I. B. (2005b). Rorschach inkblot method. In M. Maruish (Ed.), *The use of psychological testing in treatment planning and outcome evaluation* (3rd ed., Vol. 3, pp. 553–588). Mahwah, NJ: Erlbaum.

Weiner, I. B. (2006). The Rorschach inkblot method. In R. P. Archer (Ed.), *Forensic uses of clinical assessment instruments* (pp. 181–207). Mahwah, NJ: Erlbaum.

Weiner, I. B. (2007). Rorschach assessment in forensic cases. In A. Goldstein (Ed.), *Forensic psychology: Emerging topics and expanding roles* (pp. 127–153). Hoboken, NJ: Wiley.

Weiner, I. B., & Bornstein, R. J. (2009). *Principles of psychotherapy: Promoting evidence-based psychodynamic practice* (3rd ed.). Hoboken, NJ: Wiley.

Weiner, I. B., & Exner, J. E., Jr. (1991). Rorschach changes in long-term and short-term psychotherapy. *Journal of Personality Assessment, 56*, 453–465. doi:10.1207/s15327752jpa5603_7

Weiner, I. B., & Greene, R. L. (2008). *Handbook of personality assessment*. Hoboken, NJ: Wiley.

Westen, D. (1991). Social cognition and object relations. *Psychological Bulletin*, *109*, 429–455. doi:10.1037/0033-2909.109.3.429

Westen, D., Lohr, N. E., Silk, K., Gold, L., & Kerber, K. (1990). Object relations and social cognition in borderlines, major depressives, and normals: A Thematic Apperception Test analysis. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 355–364. doi:10.1037/1040-3590.2.4.355

Wood, J. M., Nezworski, G., Garb, H. N., & Lilienfeld, S. (2001). The misperception of psychopathology: Problems with the norms of the comprehensive system. *Clinical Psychology: Science and Practice, 8*, 360–373.

CHAPTER 11

ASSESSMENT OF PERSONALITY AND PSYCHOPATHOLOGY WITH SELF-REPORT INVENTORIES

James N. Butcher, Shawn Bubany, and Shawn N. Mason

The most direct and informative approach to learning about a person's problems, attributes, motivations, and resources is to ask the person directly. The self-report approach to appraising personality using questionnaire inventories has been evolving for about 100 years (Heymans & Wiersma, 1906; Woodworth, 1920). Psychologists have attempted to develop effective approaches to assessing personality information provided by the person being assessed. Although the personal disclosures of clients provide personality information not available through observation or history or background, there are many factors (e.g., assessment context, level of cooperation of the client and some demographic factors) that, if not addressed, can limit the value of self-reported symptoms and attitudes through questionnaires.

In this chapter, we provide a perspective on contemporary personality assessment methodology by describing scale development strategies and examining the factors central to making an effective, fair, and valid self-report personality measure. Several theoretical and statistical approaches to constructing self-report measures are described and compared. Influential demographic factors such as gender, ethnicity, testing circumstances, and their potential effect in personality assessment are considered. We highlight factors pertinent to the sound international adaptation of personality scales. We include a summary of several of the most widely used self-report personality measures and survey the extent of their research base that has accumulated over the past 27 years. Finally, important considerations for assuring responsible assessment applications and prospects for future developments are considered.[1]

METHODS OF CONSTRUCTING PERSONALITY QUESTIONNAIRES

Several scale construction methods have been used to develop personality questionnaires since the beginning of this movement in the early 20th century (see Butcher, 2010). These varying approaches have been used as a means of capturing personality dimensions or traits through use of a client's willingness to share personal information.

Theoretically Derived One-Dimensional Personality Inventories

The first self-report personality inventory, the Personal Data Sheet, was developed by Robert Woodworth using a rational scale development approach. He constructed items to assess potential adjustment problems for use in screening out maladjusted draftees during World War I (Woodworth, 1920). This approach to personality scale development relies on the face validity of items to ensure their utility and relevance. This item construction strategy depends on the developer's judgment in writing items that have an obvious theoretical relationship to the personality characteristics being assessed.

[1]We also note that there is another chapter on personality assessment in the Counseling Psychology section of this handbook (see Chapter 24, this volume). This latter chapter describes personality assessment used in clinical settings where psychopathology is less of a concern and individuals seeking counseling are the focus. Additionally, Samuel E. Krug has contributed a basic chapter on objective personality assessment (see Volume 1, Chapter 19, this handbook).

DOI: 10.1037/14048-011
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

Theoretically Derived Multidimensional Personality Inventories

Whereas the Personal Data Sheet involved a single dimension of adjustment–maladjustment, several subsequent, rationally derived inventories broadened the scope of assessment to include multiple dimensions. In developing their inventories, Bernreuter (1931) and Humm and Wadsworth (1934) used a similar rational/theoretical development strategy to construct items. As with Woodworth's inventory, the authors relied on their own judgment to develop items that addressed the characteristics they chose but did not empirically validate the effectiveness of the resulting scales.

Empirically Derived Psychological Test Measures

A very different approach to developing psychological assessment measures, empirically derived scales, emerged in the late 1930s and was described by Paterson, Williamson, and Schneidler (1938). They did not accept the rational scale development method that was widely used in constructing psychological measures because they thought that some items could be predictive of relevant criteria without having an apparent or obvious content connection. They believed that items needed to have proven utility before they were incorporated in the scale;.that is, items should be selected on the basis of their validity in detecting the characteristics in question.

The first personality measure that based its item selection on strictly empirical validation was the Minnesota Multiphasic Personality Inventory (MMPI), developed by Hathaway and McKinley in the 1940s (Hathaway & McKinley, 1940, 1942, 1943). The MMPI clinical scales were developed following a strategy in which items were selected for a particular scale if they empirically differentiated a well-defined clinical patient group from a normal sample. For example, items were selected for the Depression scale if they actually discriminated a group of depressed patients from a sample of nonpatients (or *normals*). Items on empirical scales may be diverse in content and not necessarily internally related (correlated) to each other. Thus, scale content can be heterogeneous. What matters most is the predictive association with the criteria, a quality not ensured in earlier personality test development. The scales developed by Hathaway and McKinley were maintained in the restandardization of the MMPI (MMPI–2) in 1989 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Butcher et al., 2001) to ensure that the research accumulated on the original test would be relevant to the revised instrument.

Personality Measures Developed by Factor Analysis

The use of factor analysis to develop personality scales was pioneered by Cattell in the 1940s (Cattell, 1946). In this approach, correlational–statistical procedures are used to ensure that the item content (derived from a large pool of potential items) selected for each scale measures a homogeneous dimension. Once the factor dimensions of the item pool are obtained, appropriate scale names summarizing the content are assigned to them. Because items for a scale are selected on the basis of intercorrelation procedures, the scales tend to be very homogeneous in the item content and narrowly defined (for more information on this technique, see Volume 1, Chapter 5, this handbook).

Sequential System of Construct-Oriented Scale Development

Jackson (1970) developed a somewhat different approach to scale construction using a modification of strict factor-analytic procedures. First, personality constructs are theoretically defined. Next, a relevant item pool to measure these potential constructs is created. Then, factor analysis is used to ensure homogeneity of constructs by ascertaining that the items are highly correlated with the total score. This factorial/dimensional strategy results in homogeneous content scales that are typically recognizable to test takers and somewhat open to response manipulation.

Recently, this strategy was applied as a method to modify the MMPI clinical scales that were originally developed following the empirical scale construction approach (Tellegen et al., 2003). Tellegen and colleagues constructed the MMPI–2 Restructured Clinical Scales (RC Scales), using a strategy modeled after Jackson's approach, to develop a set of measures that were highly homogeneous and narrow in construct meaning resulting in a set of scales that

have minimal resemblance to Hathaway and McKinley's empirically developed clinical scales (Nichols, 2006; Ranson, Nichols, Rouse, & Harrington, 2009). The developers of the resultant test—the MMPI–2—Restructured Form (MMPI–2–RF)—did not ensure equivalence with the MMPI–2. Although it uses a subset of the MMPI–2 item pool and subjects from the 1989 normative sample, it is a new measure that cannot rely on previous studies involving the MMPI or MMPI–2 for interpretative research.

Content-Based Personality Measures

An effective means of constructing personality scales involves a combination of the rational and correlational statistical approaches described earlier. The content scale developmental strategy advanced by Wiggins (1969) uses successful features of both approaches. First, items are grouped into clusters based on similarity of content and a rational connection to a theoretical construct or personality characteristic, such as somatic symptoms in the Poor Health scale and religious beliefs in the Religious Fundamentalism scale. This approach is derived from Cronbach and Meehl's (1955) construct validity approach of using a combination of content grouping and statistical refinement. Constructs such as personality traits may be used as the basis for developing the item pool for the scale. Item analysis and related procedures such as coefficient alpha can be used to ensure that the items are associated with the scale construct and guarantees that the scale has internal consistency and scale homogeneity. In some developmental projects, such as in the MMPI–2 Restandardization Project (Butcher, Graham, Williams, & Ben-Porath, 1990), the content scales are also externally validated by ensuring that meaningful correlates are obtained. This strategy, as in the Sequential System approach described previously, results in scales that are generally homogeneous in content and somewhat narrow in construct.

Item Response Theory (IRT)

A more contemporary psychometric approach, referred to as *item response theory* (IRT), has received considerable research attention in recent years, although no widely used commercial personality scales have been developed with this method. Much of the research using IRT has focused on scale refinement (Glas, 2009; see also Volume 1, Chapter 6, this handbook) and computer-adaptive measurement (see Thompson, 2009). This strategy for development or refinement involves item analysis to define statistically characteristics underlying a measured dimension. It is based on the idea that responses to test items reflect an underlying variable, such as a trait. Embretson (1996) provided a description of how test construction according to IRT differs from that of traditional scale construction procedures and proposed that these are fundamentally different from classical test theory and require a different approach to item selection. A personality scale might be developed by reference to the underlying dimension. An effective IRT approach to item refinement typically follows two specific assumptions: that the item pool should be unidimensional (i.e., that all items assess a single latent trait) and that the probabilities of the specified responses at different levels of the underlying trait can be fit adequately by some IRT model. The IRT approach has been shown to be a valuable procedure for constructing or refining personality scales (Glas, 2009; Reise & Waller, 1993).

Relative Effectiveness of Different Approaches to Scale Construction

Research on the relative performance of the rational, empirical, and factor-analytic approaches has not found any of these methods to be superior to the others. Hase and Goldberg (1967) and Burisch (1984) reported that the empirical, rational, and factor-analytic approaches were equally valid development methods. Burisch, in his replication of the Hase and Goldberg study, found that each test construction method yielded comparable results. He concluded that the deductive, rational approach might be preferable because it is a more efficient strategy and recommended that test constructors consider using the rational method of developing tests.

TEN BASIC REQUIREMENTS FOR A VALID AND EFFECTIVE PERSONALITY SCALE

Regardless of the purposes intended for the test, which developmental strategies are used in

construction, what the nature of the constructs are, or who the intended populations the test is designed to address, there are general guidelines for the test developer to consider and for the practitioner to follow in making decisions as to whether to use the scale in assessment. Several articles have been published that focus on these issues and recommendations for scale developers to consider (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Butcher, Graham, & Ben-Porath, 1995; Geisinger, 2005). Important steps in ensuring that a personality measure meets expectations of practitioners and researchers include the following:

1. The personality constructs addressed by the scale(s) need to be well defined and well grounded. The variable(s) in question should not simply be the result of a vague "theory" or some interesting group differences derived from a sample of convenience. If the construct for which the scale is developed is theory based, then clear, well-established theoretical guidelines are important. Simply referencing a theory as the source is insufficient. Clear, theoretically based hypotheses and inferences need to be formulated. The personality construct or symptom pattern should be substantiated by defined research. It is crucial that the dimension(s) addressed be well-delineated personality characteristics and not superficial or irrelevant variables or characteristics that are unrelated to personality (see Ranson et al., 2009). For example, some human characteristics such as abilities, intellectual qualities, or vague attributes are questionable constructs for personality inventories.

2. The item pool developed for the scale(s) needs to be relevant and sufficient for the personality construct being assessed. Scales for special attributes need to have appropriate item contents. Items can be developed following several models, as noted earlier. It is critical that the items cover the concepts being addressed in sufficient detail to result in a reliable measure (e.g., items developed for assessing specific problems should not be considered as appropriate for developing measures for all personality dimensions). Although the original MMPI, for example, was developed to assess psychiatric problems, the items were used for hundreds of scales (Dahlstrom, Welsh, & Dahlstrom, 1975) that covered an extreme range of constructs such as Success in Baseball, Tired Housewife, Worried Breadwinner, Effective Physician, and so forth. Any personality inventory, even one as long as the MMPI with its 566 items, has limitations. It is unlikely that the inventory provided sufficient item relevance and depth to support the broad number of scales that were created and published. In developing a personality scale, it is important to ensure that the item content is relevant and sufficient for representing the construct being addressed. Some scales, such as the NEO Personality Inventory (NEO PI; Costa & McRae, 2009), were developed for normal range personality assessment, yet they are used for addressing clinical problems. Similarly, the MCMI instruments were developed for use with patients in psychotherapy, yet they have been used in settings such as family custody or personnel screening, for which they neither were designed nor have an appropriate item set or normative base.

3. The items selected for inclusion in the scale(s) should be linguistically and psychologically appropriate and sufficient for reliable assessment. The items should be written at a level of simplicity to allow persons with low reading or comprehension levels to understand them. In addition, given that many scales are eventually translated into other languages, the items should not contain highly culturally specific content that thwarts general use or adaptation.

Moreover, the scales in an inventory need to be of sufficient length. That is, scales must contain a sufficient number of appropriate items to address the constructs being measured (Emons, Sijtsma, & Meijer, 2007). Brevity alone is not an effective criterion or goal for assessment measures. Comprehensive validity and target coverage are more appropriate goals. Hathaway (1975) cautioned against shortened versions of the MMPI, for example, because he believed that any assessment of clinical and personality problems required an extensive amount of information from the client. However, there is a trend in personality assessment that has resulted in weak or flawed measures by following the assumption that shorter or abbreviated is better. For example, some efforts have attempted to develop shortened versions of MMPI scales but have

failed to capture the reliable and valid original measures (see Butcher & Hostetler, 1990; and Dahlstrom, 1980, for a discussion of the problems encountered with shortened MMPI scales).

The MMPI–2–RF contains a number of very brief scales (four to six items) that have extremely low reliabilities for personality scales. For example, the internal consistency coefficients for the Helplessness scale (five items) were only .39 for men and .50 for women in the normative sample; the coefficients for the Behavior–Restricting Fears scale (nine items) were only .44 for men and .49 for women; and the coefficients for the Suicidal/Death Ideation scale (five items) were only .41 for men and .34 for women (Tellegen & Ben-Porath, 2008). These low reliabilities raise questions about the meanings and reproducibility of these measures for the respective constructs. Moreover, many of these new and abbreviated scales are not supported by behavioral correlates (Butcher, 2011; Greene, 2011; Tellegen & Ben-Porath, 2008).

4. The item pool for the personality scale(s) being developed should allow for fair and balanced assessment of the desired personality characteristics. Personality scales, because they often are used in making high-stakes decisions in forensic cases or personnel selection, need to be fair and balanced (Geisinger, 2005; Geisinger & Carlson, 2009). Any content that can result in discriminatory or unfair evaluations in high-stakes assessments should be eliminated, or the test should not be used. The Fake Bad Scale (FBS; renamed the Symptom Validity Scale by its publisher in 2008; Lees-Haley, English, & Glenn, 1991), developed with MMPI items, is a case in point. This scale, designed for personal injury evaluation, has been shown to discriminate against people with actual physical disability because a large percentage of the FBS items also appear on scales that measure stress or somatic problems (Butcher, Arbisi, Atlis, &McNulty, 2003; Butcher, Gass, Cumella, Kally, & Williams, 2008; Gass, Williams, Cumella, Butcher, & Kally, 2010; Williams, Butcher, Gass, Cumella, & Kally, 2009). Ben-Porath and colleagues (e.g. Ben-Porath, Graham, & Tellegen, 2009) have provided alternative views about the utility of the FBS despite the concerns described about its item pool.

5. The scale development needs to include appropriate statistical information that allows potential users to evaluate its merits. The research sample sizes need to be sufficiently large to enable the computation of reliability and validity statistics. The research sample sizes should be large enough for dividing into sub-samples to allow cross-validation of developed measures. Any new scale should have high internal consistency if the scale has been designed as a measure of a single dimension or trait. Some scales, such as empirically derived infrequency measures, do not require internal consistency because they are composed of multiple item content groupings and homogeneity of item content is less relevant. Cross-validation of measures is particularly important for empirically derived scales to ensure that chance items are eliminated.

It is important for most scales or sets of scales to have a meaningful factor structure to enable users to have a clear idea of the constructs being measured. The scale's factor structure should be examined and reported for the measure(s) to be appraised by potential users. In the case of empirically derived measures, high internal consistency is less important than test–retest reliability and external validity values.

6. The research design should include sufficient research samples that allow for the development of appropriate reference populations or norms for the scales and for cross-validation of the measures developed. Normative samples should be of sufficient size and demographic heterogeneity to allow for the development of representative norms (Geisinger & Carlson, 2009; Wasserman & Bracken, 2003). The scale development samples from both the target population and from the reference or normal populations should allow for evaluating the generalizability of the developed measures.

7. Construct validity for the underlying variables addressed by the scale should be reported. Anastasia and Urbina (1997) described construct validity as the extent to which a test measures a theoretical construct or trait. Smith and Zapolski (2009) pointed out that there are two crucial limitations to simply using criterion validity to establish the meaning of a construct: The validity results are only as good as the criteria they predicted; and the process adds little to the basic theory for the scale. In their

classic approach to the validation of constructs, Cronbach and Meehl (1955) postulated establishing a "nomological network" that involved specifying the lawful relationships between an inferred construct and other related constructs. The construct validity of a particular measure needs to incorporate a network of relationships and not simply focus on a single variable or criterion.

One approach to establishing construct validity for a new personality measure is to provide relationships, correlations, between the proposed scale and other, well-established measures on the same inventory. For example, new MMPI–2 measures need to be well described and defined by other existing measures in the same domain (see Butcher, Graham & Ben-Porath, 1995; Butcher & Tellegen, 1978, for further discussion). For new MMPI–2 measures, it is valuable to be able to understand the interscale relationships between the new or proposed scale and other well-established scales or dimensions, such as the anxiety and the repression factor dimensions, as a means of establishing the new scale's independence. Knowing the differences and similarities between new measures and existing scales could help establish the meaning(s) of the new measure or construct.

When such information is absent in the developmental presentation of new measures such as the MMPI–2 RC Scales (Tellegen et al., 2003), then subsequent research might show flaws that were not originally revealed. The test manual for the RC Scales did not provide statistical comparisons with a number of widely used MMPI–2 measures (e.g., the Content scales and PSY-5 scales); and, as it turned out, the RC Scales were found to be largely redundant measures of a number of these scales (see Greene et al., 2009; Rouse, Greene, Butcher, Nichols, & Williams, 2008). (For a discussion of the failure to provide relevant construct validity information, see Nichols, 2006; Ranson et al., 2009; and Rogers, Sewell, Harrison, & Jordan, 2006).

8. All scales, even those developed by rational or internal consistency methods, should predict observable behavior considered to represent the constructs being measured. It is crucial for any personality scale or inventory to have predictive validity. The measures need to be validated for the intended use (Geisinger, 2005). The measure should have reported validity coefficients with known criterion groups for potential test users to be able to evaluate their effectiveness at measuring what they are purported to assess.

9. The test's effectiveness needs to be explored and demonstrated. Does the scale possess sensitivity and specificity in predicting the qualities it is supposed to assess? How well the scale classifies relevant cases should be reported. It is important for tests not to be oversold for uses or populations for which the test was not intended or established through research. Practitioners should demand proof of validity and utility before tests are used to make important decisions about people.

10. A comprehensive test manual or test user's guide should be made available to potential test users. Adequate documentation with sufficiently detailed descriptions of the procedures followed in scale development to permit their replication by other researchers using different samples should be provided, and specific and clear instructions for the measures need to be provided in a test manual to guide test users in the appropriate and established test applications. The manual should enable psychologists to gain a clear overview of what the personality test measures and what are its limitations.

ESSENTIAL COMPONENTS OF PERSONALITY INVENTORIES

Several important variables can result in incomplete or faulty information in self-report personality questionnaire assessment. These factors need to be explored to ensure that their potential adverse effects on information from the test results can be minimized.

Importance of Standard Instructions

Most personality tests were developed with, and are tied to, the instructional directions used to administer items to clients. Standardized tests are expected to be equivalent across administrations (Sireci, 2005). Thus, the administration procedures need to strictly follow the administration guidelines and instructions used in the scale norming. There are clearly defined sets of instructions associated with each personality measure—these instructions are

designed to elicit cooperation and appropriate "mental sets" that encourage clients to respond in a manner that was established for the normative data collection. For test results to be reliable and valid, the scores need to be compared with those of the normative group who followed the same set of instructions.

If the instructions are altered or not carefully followed, then the client responding to the items might view the task differently and respond to the items inappropriately. Thus, different results may be obtained when compared with the normative population. For example, one psychologist chose to administer the MMPI–2 in forensic evaluations by informing the client, "Please answer the items as you felt and thought before the crime." Such "personalized" instructions invalidate the test because the standardized instructions were not followed and the normative sample was not appropriate for the client.

One contemporary test application in which test instructions and norms might deviate from the test standards and norms involves Internet administration of what were previously paper-and-pencil tests. This means of information gathering about people is different from traditional assessment and treatment and raises issues of lack of comparability (Buchanan, 2002; Meade, Michels, & Lautenschlager, 2007). For example, there has been limited research on whether the testing conditions are equivalent. Moreover, equivalence of normative databases and similarity in test-taking attitudes are lagging behind the technological expansion. Many psychological tests are interpreted by comparing the scores of a particular client to those of a known standardization group or normative sample collected under controlled conditions. Some available evidence suggests that tests administered on the Internet produce somewhat different results than those administered under standard conditions (Buchanan & Smith, 1999).

Potential for Noncredible Information as a Result of Client Response Sets

Reliance on self-report personality measures assumes that the client responded to the items in an appropriate, cooperative, and open manner and was willing to share information in the assessment. Without a generally open and cooperative self-report, personality inventory responses have no or little utility. People being evaluated in some situations (e.g., personal injury litigation, insanity pleas, or personnel screening) have strong motivations to present themselves in particular ways—to appear extremely well adjusted or to appear severely stressed with symptoms when they are not. Thus, the information gained through self-report may be low in credibility. In some situations, the most useful information in the assessment is that the client was uncooperative and produced an unclear picture of his or her adjustment. Recently, Holden (2007) reaffirmed that socially desirable responding is an important influence on self-report inventory responding and needs to be taken into consideration in personality test interpretation.

Some personality assessment measures include procedures that detect response sets whereas others do not. Contemporary personality scales address response sets differently, and psychologists need to take into consideration their proven capability of identifying deviant responding (see Arbisi, & Butcher, 2004; Berry, Sollman, Schipper, Clark, & Shandera, 2009; Pope, Butcher, & Seelen, 2006).

Gender Differences in Personality Inventory Responses

Gender differences in psychological variables have been extensively studied and debated in recent years (Guimond, 2008; Hyde, 2005). Hyde (2005) has dated interest in psychological gender differences back to the "dawn of formalized psychology" itself (p. 581). Feingold (1994) traced the history of gender difference research, noting a progression from an early emphasis on biologically deterministic views of individual difference in traits, to a contemporary focus on gender differences in cognitive abilities and social behavior. Eagly and Wood (1999) also identified an increase in the study of both personality and ability differences between men and women.

Despite significant advances in research on gender differences, there is debate as to the origins and the relative magnitude of differences and similarities across psychological variables (Eagly & Wood, 1999; Feingold, 1994; Guimond, 2008; Hyde, 2005; Kling, Hyde, Showers, & Buswell, 1999). Although

Feingold's (1994) meta-analysis found evidence of gender differences in a number of personality assessment measures, Hyde (2005) considered men and women similar on many psychological variables and suggested that most psychological gender differences are small.

Regardless of the magnitude or origin of psychological gender differences, studies continue to find evidence of differences between men and women on a number of variables. Schmit, Realo, Voracek, and Allik (2008) suggested that gender differences in personality traits tend to be larger than those in other psychological factors. Psychometricians have researched differences in the area of personality and assessment since the early 20th century and focused on deciding between combined and separate gender norms for assessment measures (Feingold, 1994). In a study comparing personality trait factors of women and men, Cattell (1948) found that, on a general emotionality scale, women's responses had more emphasis on "fearful emotionality and neurotic traits"; additionally, on a dominance scale, the women's factor emphasized attention-getting and hypochondria, whereas in the men's factor, toughness and boastful self-assertion was emphasized. Hathaway and McKinley (1942) noted significantly higher Depression scale scores for women than men on the original MMPI.

More recently, the MMPI–2 has come under scrutiny for potential gender bias in the FBS; Lees-Haley et al., 1991), which was developed to detect malingering in personal injury cases (Butcher et al., 2008; Nichols, Williams, & Greene, 2009). On this scale, women are more likely to endorse several of the somatic complaints and, thus, score higher than men (Butcher et al., 2008). The FBS test manual authors (Ben-Porath, Graham, et al., 2009), however, reached the conclusion that there is an absence of gender bias in the FBS. (For further discussion of this issue, see Gass et al., 2010; Williams et al., 2009; and responses by Ben-Porath, Greve, Bianchini, & Kaufmann, 2009.)

An important area to consider, in the role of gender in assessment, is how perceived gender roles and stereotyping affect testing outcomes. Artifact models of gender differences in assessment suggest that differences result from respondents' ideas about what are socially desirable traits depending on one's gender (Schmitt et al., 2008). Indeed, Keogh (2004) suggested that, on the Anxiety Sensitivity Index, more women than men endorse feeling worried about having a mental illness when feeling nervous because of the wide belief that women are more susceptible to certain forms of pathology. Social–structural theories similarly posit that socially pervasive gender roles and expectations result in the observed differences in behaviors and traits between men and women (Eagly & Wood, 1999).

Research in education offers additional ways of thinking about issues of fairness in test taking and assessment. It is important to note that, according to Willingham and Cole (1997), in educational testing contexts, women tend to have significantly higher test anxiety than men. When negative stereotypes (and, thus, psychological threat) are removed from a testing situation, ethnic minority students and women perform better on tests (Walton & Spencer, 2009; see also Volume 3, Chapter 27, this handbook).

Some researchers call for a consideration of context when engaging in research on gender or clinical assessment with women. Hyde (2005) noted that context can influence the magnitude and direction of gender differences, citing interactions between participants and experimenters in research studies and written instructions for examinations as potential influences on outcome. Worell and Robinson (2009) have suggested a multimodal approach, citing several ways in which understanding the broader context of an individual's life can improve clinical assessment with women. They specifically pointed to the increased rate of both threats and real experience of abuse, sexual assault, or other histories of trauma in women as important considerations. Worell and Robinson urged practitioners to be attentive to the potential for gender bias in assessment measures as well as clinician judgment. Furthermore, it is critically important to be aware of the ways in which women with multiple and intersecting identities (in terms of race, ethnicity, religion, sexual orientation, and social class) have not been represented in the majority of research studies on gender differences in the United States. (Worell & Robinson, 2009). Attention to the unique circumstances faced by these clients is a

necessary aspect of gender and culturally sensitive assessment and care.

In summary, established gender differences in personality calls attention to the need for researchers and clinicians to be aware of potential issues that can affect assessment with female clients. Personality assessment measures are frequently used to make decisions in clinical, forensic, career, educational, and other applied settings. These high-stakes assessments range from inclusion in psychiatric treatment (Worell & Robinson, 2009), access to career and educational opportunities, obtaining custody of children, and obtaining compensation for personal injury. In these contexts, it is particularly important to ensure fairness in assessment and interpretation through use of gender-specific norms. As a result of gender differences on personality test items, most personality inventory developers use gender-specific norms (e.g., see the Millon Clinical Multiaxial Inventory, third edition [MCMI–III], MMPI–2, Sixteen Personality Factor Questionnaire [16PF], and NEO PI, discussed later); however, some newly published measures do not provide gender-specific norms (e.g., the Personality Assessment Inventory [PAI] and MMPI–2–RF) although gender differences are reported in their test manuals. With respect to the MMPI–2–RF, there are clear gender differences in both responses to items on the scales (Tellegen & Ben-Porath, 2008) and responses to correlates reported to new scales (Butcher, 2011; Greene, 2011). Thus, these measures might have adverse effects for women in some high-stakes assessments.

Ethnic Considerations

The United States is a highly diverse society composed of people from multiple language and cultural backgrounds. Psychological practitioners are commonly challenged to perform assessments on clients with limited language skills and from diverse backgrounds. Psychological tests can have substantial generalization validity across cultures because many psychological disorders share many common features (discussed later). However, it is crucial for practitioners to ensure that the language used in the assessment instrument is appropriate for the client and that the client is sufficiently acculturated to the environment in which they are living (Geisinger, 2005; Hays, 2008). People who have not acculturated to the environment can appear more psychologically disturbed on tests than they are (Okazaki, Okazaki, & Sue, 2009).

Some instruments, such as the MMPI–2, have been widely translated and adapted to other cultures (Butcher, 1996; Cheung, 2009; discussed later). Psychologists in the United States often use translated versions of the MMPI–2 with non-English-speaking clients and may actually use test norms from other countries such as Mexico, China, France, and so on (see Butcher, Cabiya, Lucio, & Garrido, 2007, for a description of using the Mexican version of the MMPI–2 and Mexican norms for assessment in the United States).

Any psychological test used in mental health assessments, particularly those in which the outcome can have deleterious effects on the client (i.e., in personnel or forensic assessments), need to have a documented track record of fair and balanced assessment (Gray-Little, 2009). Important questions to ask of a psychological test in this regard are as follows:

- Were the test items developed to ensure that biased content does not disadvantage some test takers?
- Was the normative sample used to interpret the test appropriately balanced to include ethnic minorities?
- Are there alternative forms provided for those with limited English language skills? Is there research showing comparable performance of people from diverse backgrounds? Are there translated versions available for individuals with other language backgrounds? With respect to clients with physical disabilities that limit response to written items, are there other, more suitable, administration formats available? Is there an audio version for visually impaired test takers?
- Are there interpretive adjustments/limitations to be considered? For example, the L scale on the MMPI–2 can be slightly elevated among Hispanics (Zapata-Sola, Kreuch, Landers, Hoyt, & Butcher, 2009); thus, caution is needed to ensure that valid protocols are not considered invalid on the L scale alone.

The assessment of minorities with self-report inventories has received a great deal of attention in both empirical research and theoretical viewpoints (see discussions by Geisinger, 2005; Hays, 2008).

International/Cross-Cultural Adaptations

Today, psychological tests developed in one country are often adapted for use in other languages and cultures. Many reasons for this trend can be found. For example, the growth of psychology in other countries, the instantaneous and broad communication in the world through the Internet, and the fact that some instruments (e.g., MMPI–2) have a long and proven effective application after they have been adapted for cross-cultural use (see Butcher, 1996; Butcher et al., 2007; Butcher & Pancheri, 1976; Quevedo & Butcher, 2005).

One of the primary reasons for the interest in cross-cultural test adaptation is that many measures demonstrate evidence of generalization validity in other cultures because mental health problems are highly similar across cultures. A recent descriptive study of more than 17,000 patients in a worldwide study of schizophrenia from 37 different countries in Asia, Europe, South America, Africa, and the Middle East reported that, despite the inherent regional diversity in both patients and the health care systems studied, there are striking similarities in the responses of patients with schizophrenia to most measures (Karagianis et al., 2009).

Test translation and adaptation into other languages and cultures has a long history (Berry, 2002: Butcher, 1996; Butcher & Pancheri, 1976; Mirza, 1973). To obtain an equivalent measure in an adaptation program, several steps need to be taken:

Test item translation. The initial step in adaptation is to follow a stringent translation process. Items need to be carefully rendered in the target language—a situation that might require item substitution or drastic modification in order to capture the item meaning in the language of development. Mirza (1973) pointed out that personality test item translation is like translating poetry; the meaning needs to be carefully and thoughtfully captured even with alteration of the content of the items. An effective strategy in translating an inventory is to use two or more independent bilingual translators to translate the item pool. Once the translations are completed, then the translators meet as a group to decide on the most effective item translations.

Back-translation of the items. Once the item translations are agreed upon, then a possibly valuable procedure in the translation process includes a back-translation of the items. In this phase, the items are back-translated from the target language into the language of origin by a different bilingual translator. The resulting back-translation is then compared with the original version of the test items to determine whether the items have maintained equivalence in the retranslation process. Typically, 12% to 15% of the translated items require retranslation to capture the intended meaning (Butcher, 1996).

Equivalence verification through a bilingual test–retest study. After an acceptable translated version is obtained, then it is important to further determine the equivalence of the translated version by empirical verification. One effective approach to assuring test equivalence is to conduct an empirical study using a sample of bilingual individuals who complete both versions of the inventory. The test results are evaluated as one would conduct a test–retest study in the original development of the inventory (for examples of this process see Almagor & Nevo, 1996; and Sloore, Derksen, de Mey, & Hellenbosch, 1996).

Validity assessment. External validation studies of the translated inventory are important steps for assuring that the test performs acceptably in the target culture as it does in the country of origin (for examples, see Lucio, Ampudia, Duran, Leon, & Butcher, 2001; Manos, 1995; and Zapata-Sola et al., 2009).

OVERVIEW OF SELECT CLINICAL ASSESSMENT MEASURES

In this section, we summarize several of the most widely used and researched personality questionnaires to provide the reader with an illustration of the similarities and differences among self-report personality questionnaires.

The MCMI

The MCMI was developed by Theodore Millon and first published in 1977 as a measure of personality for use with clients in psychotherapy. Now in its third edition, the MCMI–III is considered the personality questionnaire most closely coordinated with the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994), although there is a current debate as to whether the PAI or the Personality Diagnostic Questionnaire—4 are better descriptors of antisocial personality disorder (see Guy, Poythress, Douglas, Skeem, & Edens, 2008). Millon sought to create an organized measure to assess personality pathology, which did not exist in other personality inventories at the time. Subsequent editions of the MCMI were published to stay current with the *DSM*; the second edition of the MCMI was released after publication of the third edition of the *DSM*, and the MCMI–III followed the publication of the *DSM–IV*.

The MCMI–III has 175 true–false self-report items. The inventory assesses 14 *DSM–IV* personality disorders (11 moderate, 3 severe) and 10 *DSM–IV* Axis I clinical syndromes (7 moderate, 3 severe) as well as 4 correction scales (3 modifying indices, 1 validity index). The MCMI–III has incorporated a new scale known as the Grossman Facet Scale, which specifies patient scores on clinical domains such as problematic interpersonal conduct and expressive behaviors. This scale allows clinicians to develop a better understanding of particular realms of problematic client functioning. These components are measured by ordinal scales that quantify how much and how well respondents match the constructs being assessed. On the personality disorder scales, items were divided into one group representing core features of personality that are unique to that disorder and another representing features peripheral and likely to be shared with one or more similar personality disorders (Strack & Millon, 2007).

Validation of the MCMI was carried out in three steps. The first step was theoretical–substantive validation, which emphasizes the importance and utility of a theoretical framework for test development. The second step was internal–structural validation. Because the MCMI uses a "polythetic" model, internal–structural validation stresses internal scale consistency but not scale independence, the idea being that many personality disorder scales overlap and are correlated. The final step was external–criterion validation. In this step, items are administered to one group of subjects who possess the trait to be measured, and one group who does not. Any items that show significant statistical differentiation between the criterion group (those who possess the trait) and the control group are externally valid. In this stage of validation, items are evaluated by their ability to discriminate between clinical groups (rather than against a normal sample). The norms used the MCMI–III include adult inpatient and outpatient clinical samples as well as an inmate correctional sample rather than "normal" individuals as do most personality tests. The MCMI–III utilizes separate gender norms (Groth-Marnat, 2009), and statistically significant gender differences were found on some MCMI scales (e.g., Antisocial/Aggressive, Somatoform, and Major Depression).

The MMPI–2

The MMPI was published in the 1940s to assess mental health problems in psychiatric and medical settings (Hathaway & McKinley, 1940). Within a few years, it became a standard personality instrument in a wide variety of settings. The MMPI underwent a major revision in the 1980s, resulting in two forms of the test—an adult version, the MMPI–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), and an adolescent form, the MMPI–A (Butcher et al., 1992). The MMPI–2 is a 567-item inventory composed of symptoms, beliefs, and attitudes in adults.

The MMPI–2 comprises scales that were developed according to multiple scale construction approaches. The clinical scales were developed to assess clinical symptom clusters following an empirical scale development strategy (Hathaway & McKinley, 1943); the Welsh Anxiety and Repression scales were developed by factor analysis, and the MMPI–2 content scales were developed by a content/empirical validation strategy. The norms for the MMPI–2 (N = 1,138 male and 1,462 female participants) were obtained by random-sampling participants

from diverse ethnic samples across the United States (Butcher et al., 2001). The normative *T* scores were developed separately by gender on the basis of response differences between men and women similar to those Hathaway and McKinley (1940) had previously found with the original MMPI. The scales include the following.

Validity scales. Several validity scales have been developed to evaluate the client's approach to the test. The L scale is a measure of the client's reluctance to acknowledge faults or problems. The K scale assesses test defensiveness or the tendency to minimize problems; the F scale, Back F, and Fp scales assess the tendency of some people to exaggerate problems or claim excessive symptoms; and two scales address inconsistent responding (the True Response Inconsistency Scale and the Variable Response Inconsistency Scale).

Clinical scales. The following clinical scales were developed to differentiate patients from nonpatients empirically: Hypochondriasis, Depression, Hysteria, Psychopathic Deviate, Paranoia, Psychasthenia, Schizophrenia, and Mania. In addition, two other scales were included on the clinical profile to address issues of sex role identification, the Masculinity/Femininity scale and the Social Introversion and Extraversion.

Content-based scales. The content scales are homogeneous item groups that address unitary themes and represent clear communication about problems to the practitioner. There are 15 content scales measuring different symptom areas and problems. Examples include the Antisocial Practices, Bizarre Mentation, and Family Problems scales.

Special scales. Several supplemental scales have been developed to assess specific problems. For example, the MacAndrew Addiction Scale measures the potential to develop problems of addiction, and the Addiction Potential Scale assesses whether the individual acknowledges having problems with drugs or alcohol. The Marital Distress Scale assesses clients' attitudes toward their marital relationship.

MMPI–2–RF. The MMPI–2–RF, developed from a portion of the MMPI–2 items and using the original MMPI–2 norms, was published by Tellegen and Ben-Porath (2008). This 338-item instrument does not contain the original MMPI clinical scales or the MMPI-2 content scales but is based on a new set of measures, the RC Scales (Tellegen et al., 2003), which have been highly criticized (Binford & Liljequist, 2008; Butcher, Hamilton, Rouse, & Cumella, 2006; Gordon, 2006; Nichols, 2006; Ranson et al., 2009; Rogers, Sewell, Harrison, & Jordan, 2006; Rouse et al., 2008; Simms, Casillas, Clark, Watson, & Doebbeling, 2005; Wallace & Liljequist, 2005).

Although the RC Scales were included as the core measures for MMPI–2–RF, Ben-Porath and Tellegen (2008) pointed out that "the RC scales were never thought to be sufficient for a comprehensive MMPI–2-based assessment of clinically relevant attributes" (p. 5). Consequently, they included a number of additional short measures in an effort to address other problem areas. Most of the additional scales have not been clearly described or empirically evaluated by independent researchers and have unknown or nondistinct correlates. The MMPI–2–RF, unlike other versions of the MMPI, used nongendered *T* scores rather than the gender-based *T* scores that control for gender differences found on personality test items.

Edwards Personal Preference Schedule (EPPS)

The EPPS (Edwards, 1954, 1959) is a self-report measure designed to assess 15 manifest needs hypothesized by H. A. Murray (1938): achievement, deference, order, exhibition, autonomy, affiliation, intraception, succorance, dominance, abasement, nurturance, change, endurance, heterosexuality, and aggression. The EPPS has been widely used in research and applied settings.

The EPPS scales were constructed using a rational approach. The standard version of the EPPS is an ipsative measure; however, a normative version was created with Likert-type response options. Ipsative scale scores of the EPPS standard version reflect the extent to which each need is more or less characteristic of the individual than the other needs, rather than indicating the extent to which one need is more or less characteristic than the mean of a population (i.e., normative). In contrast to measures with

rating scales or true–false response options, the EPPS standard version presents pairs of 135 distinct statements that reflect each of the needs. The EPPS directs respondents to make a "forced choice" of the statement from each pair that most represents themselves. Notably, the EPPS was designed to minimize the effects of social desirability by pairing statements independently judged to be equivalent in social desirability. The EPPS consists of a total of 225 items that include 210 distinct pairs with statements representing one of the 15 needs/motives and 15 pairs of statements that are repeated to determine the consistency of responses. Each item is scored on two of the 15 scales.

Normative scores have been reported for a college student sample of 760 males and 749 females and for an adult sample of 4,031 men and 4,932 women (Edwards, 1959). The college student norms have been criticized as inaccurately describing the college student population (Thorson & Powell, 1992). Examination of gender differences with college students indicated findings consistent with gender stereotypes. Specifically, males reported higher achievement, autonomy, dominance, heterosexuality, and aggression scores on average than females, whereas females reported higher affiliation, abasement, and nurturance scores on average than males (Edwards, 1959).

Edwards (1959) reported that the EPPS demonstrated acceptable levels of split-half and test–retest reliability. The structure of the EPPS has been generally supported by factor analysis (e.g., Levonian, Comrey, Levy, & Procter, 1959). Convergent validity of EPPS scores has been supported by correlations with scores from other personality instruments such as the NEO PI (e.g., Piedmont, McCrae, & Costa, 1992). EPPS scores have been reported to discriminate between overachieving and underachieving groups of college students (Gebhart & Hoyt, 1958). On the basis of empirical findings, researchers have called into question the ability of the EPPS to prevent faking and social desirability effects successfully (e.g., Dicken, 1959).

The NEO PI

The NEO Personality Inventory—Revised (NEO PI–R; Costa & McCrae, 1992) is a widely used measure of the five-factor model (FFM) of personality traits (i.e., Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness) as well as 30 specific traits (i.e., facets) that hierarchically make up the FFM traits. The NEO PI–R consists of 240 items with six items measuring each of the 30 specific traits and a 5-point Likert-type response scale ranging from 1 (*strongly agree*) to 5 (*strongly disagree*).

The original NEO Inventory (NEO; McCrae & Costa, 1983) was designed to measure Neuroticism, Extraversion, and Openness to Experience personality traits. The NEO was revised to create versions known as the NEO PI (Costa & McCrae, 1985) and NEO FFM (Costa & McCrae, 1985), which additionally measured Agreeableness and Conscientiousness. The NEO FFI (Costa & McCrae, 1992) consists of 60 items that are scored to reflect the FFM traits but not the facets. A rational and empirical approach was used to construct the NEO, NEO PI, and NEO PI–R scales by first generating a pool of items based on theoretical conceptualization of the traits. Factor analyses of item responses were then used to determine items that best captured the traits. The NEO PI–3 (McCrae, Costa, & Martin, 2005), a revision of the NEO PI–R, was created by modifying language used in 48 items of the NEO PI–R to address a wider population of adolescents and adults.

Separate forms for self-report and observer ratings of males and females have been developed for the NEO PI–R. The norm sample consists of 500 males and 500 females aged 21 and older selected to represent the United States census with regard to age and race. Standardized scores are given in the form of *T* scores separately for males and females. The validity of an individual's responses is assessed by three items asking if the respondent answered honestly and accurately to all items and by determining the number of missing responses, "agree" and "disagree" responses, and strings of identical responses. The NEO PI–R does not include scales measuring social desirability or inconsistency. Gender differences with NEO PI–R scores and samples from 26 cultures have been found to be moderate in magnitude (i.e., less than 0.5 *SD*) and consistent with gender stereotypes (Costa, Terracciano, &

McCrae, 2001). Females were reported to have higher negative affectivity, submissiveness, and nurturance scores, whereas males were found to have greater assertiveness scores. In addition, gender differences were found to be greater in Western cultures. The NEO PI–R has been translated into multiple languages including German, Portuguese, Hebrew, Chinese, Korean, and Japanese.

The structure of NEO PI–R scale scores has been supported by factor-analytic studies with groups of males, females, adults aged 21 to 29, adults aged 30 to 65, White individuals, people of color, and participants in several different countries including Germany, Portugal, China, Korea, and Japan as well as with self, peer, and spouse ratings (Costa & McCrae, 1995; McCrae & Costa, 1997). Convergent and discriminant validity of FFM trait and facet scores have been supported by correlations with a number of other measures of personality (Costa & McCrae, 1995). Alpha coefficients have been reported for specific facet scales to range from .56 (tender minded) to .81 (depression; Costa et al., 1991) and for FFM scales ranging from .73 (agreeableness) to .93 (neuroticism; Costa & McCrae, 1988).

The PAI

Similar to the original MMPI, the PAI by Morey (1991) was designed to assess contemporary diagnostic syndromes. The test development followed Loevinger's (1957) construct validity approach and Jackson's (1970) scale development strategy. The scales on the PAI include the following: four validity scales (Inconsistency, Infrequency, Negative Impression, and Positive Impression); 11 clinical scales (Somatic Complaints, Anxiety, Depression, Mania, Paranoia, etc.); five scales that address treatment issues (Aggression, Suicidal Ideation, Stress, Nonsupport, and Treatment Rejection); and two interpersonal scales (Dominance and Warmth).

The PAI was normed on 1,000 adults from a stratified sample and compared with 1,246 clinical patients and 1,051 college students. The PAI uses a 4-point Likert-type response format rather than a true–false response approach as many other personality scales use. Although gender differences were found for the scales (some exceeding the standard error of measurement), nongendered *T* scores are used on the test to compare clients with the test normative sample. The PAI uses normative *T* scores with an $M = 50$ and $SD = 10$. A short form of the PAI comprises the first 160 items, but limited information is available for this version. Moreover, the test manual cautions test users not to use the short form of the PAI for important decisions. Validity studies were reported in the test manual, and the relationships with other personality scales, such as the MMPI, are provided.

The 16PF

The 16PF was developed by Cattell in 1946. Its fifth edition contains 185 multiple-choice items that measure 16 bipolar dimensions of personality (warmth, reasoning, emotional stability, dominance, liveliness, rule consciousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehension, openness to change, self-reliance, perfectionism, tension), five global factors (extraversion, anxiety, tough-mindedness, independence, self-control), and three validity scales (impression management, infrequency, acquiescence). The norms were derived from a stratified random sample ($N = 10{,}261$) based on the 2000 U.S. census. The 16PF is intended for use with normal populations, such as in personnel screening. Significant gender differences have been found in the personality items on the 16PF scales; thus, separate gender norms are used for each scale (Cattell, Eber, & Tatsuoka, 1970).

The 16PF is based on a factor-analytic model of personality. Cattell originally applied factor analysis to Allport's extensive list of traits and identified a smaller number of stable "source traits" that became the basis for the 16PF (Ewen, 2003). Cattell believed that personality traits influenced how people behaved and thus designed the 16PF for use in clinical and counseling environments as well as vocational and personnel assessment, couples counseling, and educational psychology, among other applications.

The 16PF has been the subject of a substantial number of research studies. Because the questionnaire is derived from factor analysis, evidence of construct validity has been garnered from research confirming the factor structure of each

construct. Cattell subjected the 16PF to three forms of consistency—reliability, homogeneity, and transferability—across different populations (Cattell et al., 1970). Three types of validity evidence are reported: Direct concept evidence describes how a given scale correlates with the source trait it attempts to measure; circumstantial concept evidence is assessed in the 16PF by using the correlation of a given scale with the other scales; and concrete evidence is offered as descriptions of the correlation between a given scale and performance in specific areas, such as school achievement (Cattell et al., 1970).

SURVEY OF CLINICAL INVENTORY RESEARCH

To obtain a perspective on the extent and direction of clinical personality testing, we conducted a survey of the research publications of the most widely used measures over the past 27 years. The use of personality instruments in research can be observed by examining trends in the number of publications and dissertations concerning specific personality inventories over time. In doing so, the database search tools of PsycINFO and Social Science Citation Index (SSCI) were used to determine the number of yearly publications and dissertations concerning widely used assessment instruments between the years 1985 and 2011. Instruments included in the searches were the EPPS, MCMI, MMPI, MMPI–2, MMPI–A, NEO PI, PAI, and 16PF.

Together, PsycINFO and SSCI cover sources that include peer-reviewed journal articles, non-peer-reviewed journal articles, book chapters, non-English publications, and dissertations across psychology and social science literatures. PsycINFO and SSCI identify documents in which designated search terms are present in the title, abstract, and key terms. One search per instrument per database was conducted because of a nearly complete overlap between searches with multiple search terms (e.g., *Minnesota Multiphasic Personality Inventory* and *MMPI*). The PsycINFO features of "mapping to a subject heading" and "auto-explode" were available and used for searches of the MMPI, MMPI–2, MMPI–A, MCMI, MCMI–II, MCMI–III, NEO PI, NEO PI–R, 16PF, and EPPS to increase accuracy of data. Together, these features incorporate specific search terms that correspond with a broader term (i.e., subject heading) within a single search. For example, using this function finds all documents regarding the MMPI, MMPI–2, and MMPI–A with one search under the main search term of MMPI. For other searches, search terms were selected with the intention of maximizing coverage while minimizing false hits.

Specifically, acronyms were used as search terms for instruments with widely known acronyms to include the MMPI, MCMI, 16PF, and NEO PI. Using the search term *MMPI* captured documents concerning the MMPI, MMPI–2, and MMPI–A. Likewise, using the term *MCMI* led to documents concerning the MCMI, MCMI–II, and MCMI–III, and using *NEO PI* captured articles for both the NEO PI and NEO PI–R. Publications and dissertations indicated by each search were then inspected to increase accuracy. Yearly publications and dissertations for each instrument were first tallied using PsycINFO. The SSCI was then used to tally additional yearly publications for each of the instruments. Overlap in PsycINFO and SSCI searches was avoided by determining whether the source of a publication indicated by SSCI is among the PsycINFO sources.

From these methods, a plot of yearly publications depicts trends over the 27-year time period (see Figure 11.1). Most notable is the general rise of publications for the MMPI, MCMI, and 16PF from 1985 to 1995 and a decline thereafter. The rise in publications that occurred between the mid-1980s and mid-1990s corresponds with the release of the MCMI in 1983, the NEO PI in 1985, the MCMI–II in 1987, the MMPI–2 in 1989, the MMPI–A in 1992, and the NEO PI–R in 1992.

To obtain a clearer picture of the drop in research publications on self-report inventories, we also examined the publication history of the Rorschach technique, the most widely used and researched performance-based instrument (see Chapter 10, this volume). We found that the trend in the number of publications regarding the Rorschach parallels the general trends observed for the MMPI, MCMI, and 16PF. Specifically, the number of publications for the Rorschach generally rose between 1985 and 1995 and then declined.

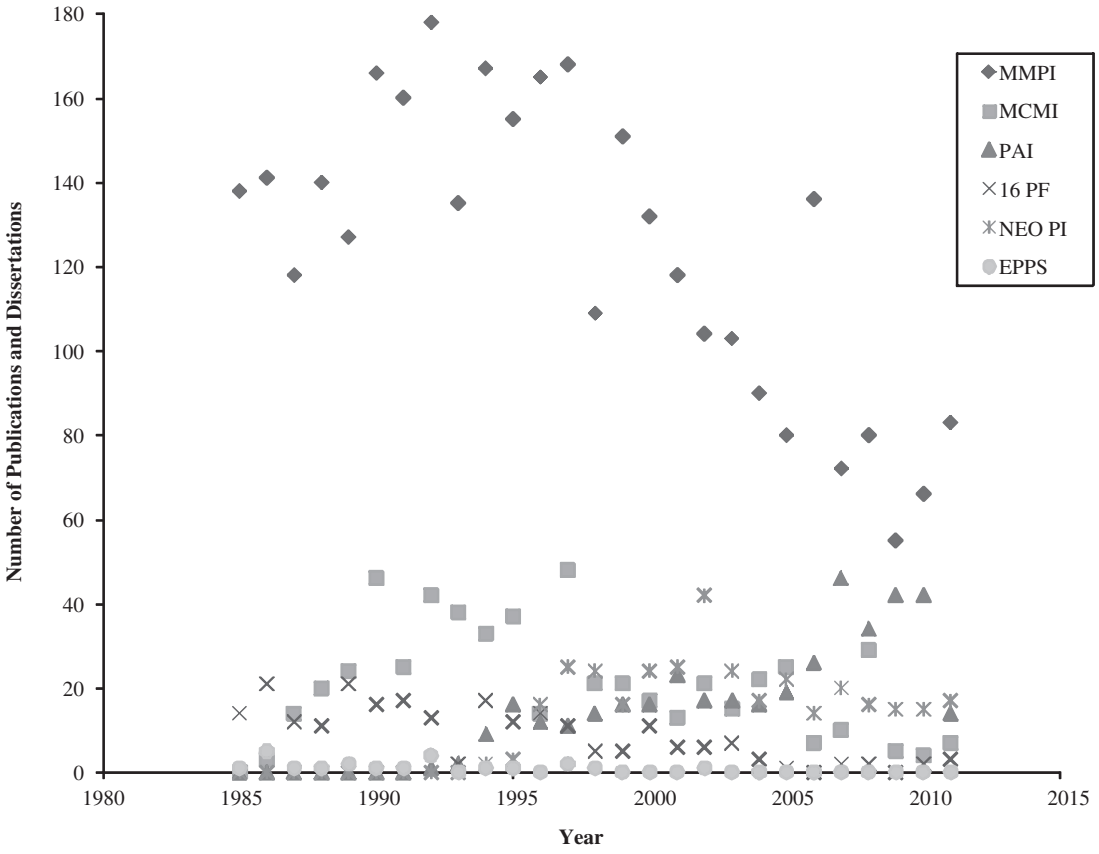


FIGURE 11.1. Number of publications and doctoral dissertations about the most widely used personality inventories over the past 27 years: Minnesota Multiphasic Personality Inventory (MMPI), Millon Clinical Multiaxial Inventory (MCMI), Personality Assessment Inventory (PAI), Sixteen Personality Factor Questionnaire (16 PF), NEO Personality Inventory (NEO PI), and Edwards Personal Preference Schedule (EPPS).

An examination of the publication trends for personality inventories shows a sharp reduction in personality assessment research over the past decade, despite the introduction of new measures. In particular, the general trend and the decline in research concerning the MCMI, despite the introduction of the MCMI–III and the introduction of the Restructured Clinical scales and the MMPI–2–RF, suggests that factors aside from fewer measures being released account for the downward trend. What factors might account for this reduction in personality research with self-report inventories in recent years? A number of contemporary situations could have contributed to this reduction in research:

- There has been a comparable reduction in the number of doctoral dissertations that focus upon self-report personality inventories. This diminishing number of graduate students who focus upon personality and or personality assessment could result from an overall lowering enrollment in research or assessment oriented programs or a modification in research goals different from personality assessment.
- New developments with other assessment approaches to research on mental health problems are receiving considerable attention in research in abnormal psychology. For example, studies to evaluate brain processes in psychiatric patients with functional magnetic imagery techniques has expanded (MacDonald & Jones, 2009).
- The reduction in research for some instruments could result from the fact that there are more personality assessment inventories available for psychologists to use in personality research today. For example, the PAI was designed to measure

many of the characteristics addressed by the MMPI scales. However, the overall reduction in published articles appears to affect most instruments.
- There are recent controversies surrounding some measures such as the MMPI–2–RF and the Rorschach (see Butcher, 2010) that could have a disparaging effect on including instruments in assessment studies.
- Concerns over personality tests not delivering valid and useful information as promised or controversial applications such as using the FBS to deny injury claims could have tainted such measures.
- The fact that insurance reimbursements for psychological assessments have diminished might make personality assessment studies more difficult to accomplish, because testing results are more limited. In addition, many psychological clinical practices have changed their focus toward forensic assessment and are less amenable to conducting clinical research.
- This reduction in research on self-report questionnaires might, in part, result from an overall reduction in funding resources for research in mental health problems with personality inventories.

These possible reasons notwithstanding, the decline in publications raises important concerns about the extent to which personality assessment research is adequately meeting the ongoing and expanding needs of practitioners or being informed by advances in basic research. We have attempted to obtain a perspective of self-report inventories use over the past 27 years. Keep in mind that this survey is not a test-use survey but a summary of published research articles, books, book chapters, and dissertations. We do not have access to information pertaining to whether test use has stayed the same, declined, or increased.

CHALLENGES FOR PERSONALITY ASSESSMENT

The diverse field of self-report personality inventory assessment has almost a century of successful development and application in psychology. An extensive array of instruments has emerged to address both personality characteristics and clinical symptom patterns. The profession of psychology has grown and has been broadly accepted by society, in part, as a result of the utility of psychological assessment.

The availability of diverse self-report instruments ensures that this aspect of patient symptom descriptions and behavior that has been a valuable component to patient understanding is likely to continue into the future, although other methods of data collection and patient problem description are likely to expand. Self-report instruments have a proven value in assessing clients in psychological evaluations if the instruments are well conceived and psychometrically robust and if the response context is clearly understood. Assessment psychologists who rely on self-report instruments face a number of challenges, however, if the field is going to develop further, maintain its broad public acceptance, and contribute personality measures that are effective and valid for the stated purposes of the instrument. Challenges include the following:

- If new or redeveloped psychological tests are to be acceptable to the field, there needs to be considerable attention paid to ensuring that the instrument meets the highest standards. The bar for publication of revised or new personality tests needs to be raised to ensure that tests deliver as promised. Before psychological tests are distributed, their psychometric properties need to be sufficiently explored and verified. Psychometric properties such as validity and reliability need to be amply demonstrated.
- Test applications and limitations need to be carefully specified, evaluated, and clearly described to ensure that test usage is appropriate.
- Personality scales should take into consideration demographic factors such as gender and ethnicity, which have been shown to have crucial influences on item responses. When personality differences are demonstrated, then procedures (e.g., gender-specific norms) can be promulgated to allow for more accurate comparisons, particularly in high-stakes assessments.
- Avoidance of harm in psychological test use is critical. If a psychological test is vulnerable to misapplication or its use could result in potential harm to clients, then such procedures need to be clearly disclosed to potential users to avoid

misuse. Using tests that can result in harm to clients would likely create suspicion of and disrespect for personality assessment in the community.

SUMMARY

This chapter provides a perspective on contemporary personality assessment methodology and instruments by describing the history and use of five scale development strategies that have been used to construct or refine self-report personality measures. We also examined factors that are important to consider in developing effective, fair, and valid self-report personality measures. These factors include instructional sets, client response sets, and demographic factors such as gender and ethnicity and their potential effect on personality test results. Because of the expansion of psychology (particularly, personality) assessment instruments, factors pertinent to sound international adaptation of personality scales were described. Six of the most widely used personality measures were described, and a survey of personality test use over the past 27 years was provided. The recent changes in personality test research were discussed. Finally, important considerations for assuring responsible assessment applications and prospects for future developments were considered.

References

Almagor, M., & Nevo, B. (1996). The MMPI–2: Translation and first steps in its adaptation. In J. N. Butcher (Ed.), *International adaptations of the MMPI–2* (pp. 487–505). Minneapolis: University of Minnesota Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

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

Anastasia, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Arbisi, P. A., & Butcher, J. N. (2004). Psychometric perspectives on detection of malingering of pain: Use of the Minnesota Multiphasic Personality lnventory—2. *Clinical Journal of Pain, 20*, 383–391. doi:10.1097/00002508-200411000-00002

Ben-Porath, Y. S., Graham, J. R., & Tellegen, A. (2009). *The MMPI–2 Symptom Validity (FBS) Scale development, research findings, and interpretive recommendations.* Minneapolis: University of Minnesota Press.

Ben-Porath, Y. S., Greve, K. W., Bianchini, K. J., & Kaufmann, P. M. (2009). The MMPI–2 Symptom Validity Scale (FBS) is an empirically validated measure of over-reporting in personal injury litigants and claimants: Reply to Butcher et al. *Psychological Injury and Law, 2,* 62–85, 2008. doi:10.1007/s12207-009-9037-4

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI–2–RF: Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Bernreuter, R. G. (1931). *The personality inventory*. Palo Alto, CA: Consulting Psychologists Press.

Berry, D. T. R., Sollman, M. J., Schipper, L. J., Clark, J. A., & Shandera, A. L. (2009). Assessment of feigned psychological symptoms. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 613–642). New York, NY: Oxford University Press.

Berry, J. W. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). New York, NY: Cambridge University Press.

Binford, A., & Liljequist, L. (2008). Behavioral correlates of selected MMPI–2 Clinical, Content, and Restructured Clinical Scales. *Journal of Personality Assessment, 90*, 608–614. doi:10.1080/00223890 802388657

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33*, 148–154. doi:10.1037/0735-7028. 33.2.148

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125–144. doi:10.1348/000712699161189

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*, 214–227. doi:10.1037/0003-066X. 39.3.214

Butcher, J. N. (1996). *International adaptations of the MMPI–2*. Minneapolis: University of Minnesota Press.

Butcher, J. N. (2010). Personality assessment from the 19th to the early 21st century: Past achievements and contemporary challenges. *Annual Review of Clinical Psychology, 6*, 1–20. doi:10.1146/annurev. clinpsy.121208.131420

Butcher, J. N. (2011). *Beginner's guide to the MMPI–2* (3rd ed.). Washington, DC: American Psychological Association.

Butcher, J. N., Arbisi, P. A., Atlis, M., & McNulty, J. (2003). The construct validity of the Lees-Haley Fake

Bad Scale (FBS): Does this scale measure malingering and feigned emotional distress? *Archives of Clinical Neuropsychiatry, 18*, 473–485.

Butcher, J. N., Cabiya, J., Lucio, E. M., & Garrido, M. (2007). *Assessing Hispanic clients using the MMPI-2 and MMPI-A.* Washington, DC: American Psychological Association. doi:10.1037/11585-000

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory—2 (MMPI–2): Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Gass, C. S., Cumella, E., Kally, Z., & Williams, C. L. (2008). Potential for bias in MMPI-2 assessments using the Fake Bad Scale (FBS). *Psychological Injury and Law, 1*, 191–209. doi:10.1007/s12207-007-9002-z

Butcher, J. N., Graham, J. R., & Ben-Porath, Y. S. (1995). Methodological problems and issues in MMPI/MMPI–2/MMPI–A research. *Psychological Assessment, 7*, 320–329. doi:10.1037/1040-3590.7.3.320

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & Dahlstrom, M. W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring* (rev. ed.). Minneapolis: University of Minnesota Press.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI–2 content scales.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Hamilton, C. K., Rouse, S. V., & Cumella, E. J. (2006). The deconstruction of the Hy scale of MMPI–2: Failure of RC3 in measuring somatic symptom expression. *Journal of Personality Assessment, 87*, 186–192. doi:10.1207/s15327752jpa8702_08

Butcher, J. N., & Hostetler, K. (1990). Abbreviating MMPI item administration: What can be learned from the MMPI for the MMPI–2. *Psychological Assessment, 2*, 12–21.

Butcher, J. N., & Pancheri, P. (1976). *Handbook of cross-national MMPI research.* Minneapolis: University of Minnesota Press.

Butcher, J. N., & Tellegen, A. (1978). MMPI research: Methodological problems and some current issues. *Journal of Consulting and Clinical Psychology, 46*, 620–628. doi:10.1037/0022-006X.46.4.620

Butcher, J. N., Williams, C. L., Graham, J. R., Tellegen, A., Ben-Porath, Y. S., Archer, R. P., & Kaemmer, B. (1992). *Manual for administration, scoring, and interpretation of the Minnesota Multiphasic Personality Inventory for Adolescents: MMPI–A.* Minneapolis: University of Minnesota Press.

Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers-on-Hudson, NY: World Book Company.

Cattell, R. B. (1948). The primary personality factors in women compared with those in men. *British Journal of Psychology, 1*, 114–130.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *The handbook for the Sixteen Personality Factor (16PF) questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

Cheung, F. M. (2009). The cultural perspective in personality assessment. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 44–56). New York, NY: Oxford University Press.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Manual for the Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI)*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21–50. doi:10.1207/s15327752jpa6401_2

Costa, P. T., Jr., & McCrae, R. R. (2009). The Five-Factor Model and the NEO Inventories. In J. N. Butcher (Ed.), *Oxford Handbook of Personality Assessment* (pp. 299–322). New York, NY: Oxford University Press.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322–331. doi:10.1037/0022-3514.81.2.322

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Dahlstrom, W. G. (1980). Altered forms of the MMPI. In W. G. Dahlstrom & L. E. Dahlstrom (Eds.), *Basic readings on the MMPI* (pp. 386–393). Minneapolis: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook*: *Vol. II. Research applications*. Minneapolis: University of Minnesota Press.

Dicken, C. F. (1959). Simulated patterns on the Edwards Personal Preference Schedule. *Journal of Applied Psychology, 43*, 372–378. doi:10.1037/h0044779

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist, 54*, 408–423. doi:10.1037/0003-066X.54.6.408

Edwards, A. L. (1954). *Personal preference schedule.* New York, NY: Psychological Corporation.

Edwards, A. L. (1959). *Manual for the Edwards Personal Preference Schedule* (rev. ed.). New York, NY: Psychological Corporation.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341–349. doi:10.1037/1040-3590.8.4.341

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105–120. doi:10.1037/1082-989X.12.1.105

Ewen, R. B. (2003). *An introduction to theories of personality* (6th ed.). Mahwah, NJ: Erlbaum.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*, 429–456. doi:10.1037/0033-2909.116.3.429

Gass, C. S., Williams, C. L., Cumella, E., Butcher, J. N., & Kally, Z. (2010). Ambiguous measures of unknown constructs: The MMPI–2 Fake Bad Scale (aka Symptom Validity Scale, FBS, FBS-r). *Psychological Injury and the Law* (Published online: 22 January, 2010) DOI 10.1007/s 12207–009-9063–2.

Gebhart, G. G., & Hoyt, D. P. (1958). Personality needs of under- and overachieving freshmen. *Journal of Applied Psychology, 42*, 125–128. doi:10.1037/h0040603

Geisinger, K. F. (2005). The testing industry, ethnic minorities, and individuals with disabilities. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 187–203). Mahwah, NJ: Erlbaum.

Geisinger, K. F., & Carlson, J. F. (2009). Standards and standardization. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 99–111). New York, NY: Oxford University Press.

Glas, C. A. W. (2009). What IRT can and cannot do. *Measurement: Interdisciplinary Research and Perspectives, 7*, 91–93. doi:10.1080/15366360903117020

Gordon, R. M. (2006). False assumptions about psychopathology, hysteria and the MMPI–2 Restructured Clinical Scales. *Psychological Reports, 98*, 870–872. doi:10.2466/pr0.98.3.870-872

Gray-Little, B. (2009). The assessment of psychopathology in racial and ethnic minorities. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 396–414). New York, NY: Oxford University Press.

Greene, R. L. (2011). *The MMPI–2/MMPI–2–RF: An interpretive manual*. Boston, MA: Allyn & Bacon.

Greene, R. L., Rouse, S. V., Butcher, J. N., Nichols, D. S., & Williams, C. L. (2009). The MMPI–2 Restructured Clinical (RC) Scales and redundancy: Response to Tellegen, Ben-Porath, and Sellbom. *Journal of Personality Assessment, 91*, 222–226. doi:10.1080/00223890902800825

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Guimond, S. (2008). Psychological similarities and differences between women and men across cultures. *Social and Personality Psychology Compass, 2*, 494–510. doi:10.1111/j.1751-9004.2007.00036.x

Guy, L. S., Poythress, N. G., Douglas, W., Skeem, J. L., & Edens, J. F. (2008). Which is better—Self-report or interview based assessment of antisocial personality disorder? *Psychological Assessment, 20*, 47–54. doi:10.1037/1040-3590.20.1.47

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies for constructing personality inventory scales. *Psychological Bulletin*, *67*, 231–248. doi:10.1037/h0024421

Hathaway, S. R. (1975, February). *Comment on MMPI abbreviated forms*. In *Who owns test items? Present confusions and anxieties about 1984*. Symposium on Recent Developments in the Use of the MMPI, St. Petersburg, FL.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology: Interdisciplinary and Applied, 10*, 249–254. doi:10.1080/00223980.1940.9917000

Hathaway, S. R., & McKinley, J. C. (1942). A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *Journal of Psychology, 14*, 73–84. doi:10.1080/00223980.1942.9917111

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota multiphasic personality schedule*. Minneapolis, MN: University of Minnesota Press.

Hays, P. A. (2008). *Addressing cultural complexities in practice: Assessment, diagnosis, and therapy* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/11650-000

Heymans, G., & Wiersma, E. (1906). Beiträge zur speziellen Psychologie auf Grund einer Massenunterschung [Special contributions to psychology from a large-scale investigation]. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie, 43*, 81–127.

Holden, R. R. (2007). Socially desirable responding does moderate scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 39*, 184–201. doi:10.1037/cjbs2007015

Humm, D. G., & Wadsworth, G. W., Jr. (1934). The Humm–Wadsworth Temperament Scale: Preliminary report. *Personnel Journal, 12*, 314–323.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581–592. doi:10.1037/0003-066X.60.6.581

Jackson, D. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 61–96). New York, NY: Academic Press.

Karagianis, J., Novick, D., Pecenak, J., Haro, J. M., Dossenback, M., Treuer, T., & Lowery, A. J. (2009). The Worldwide-Schizophrenia Outpatient Health Outcomes (W-SOHO): Baseline characteristics of pan-regional observational data from more than 17,000 patients. *International Journal of Clinical Practice, 63*, 1578–1588. doi:10.1111/j.1742-1241.2009.02191.x

Keogh, E. (2004). Investigating invariance in the factorial structure of the anxiety sensitivity index across adult men and women. *Journal of Personality Assessment, 83*, 153–160. doi:10.1207/s15327752jpa8302_09

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin*, *125*, 470–500. doi:10.1037/0033-2909.125.4.470

Lees-Haley, P. R., English, L. T., & Glenn, W. J. (1991). A fake bad scale on the MMPI–2 for personal injury claimants. *Psychological Reports, 68*, 203–210. doi:10.2466/pr0.1991.68.1.203

Levonian, E., Comrey, A., Levy, W., & Procter, D. (1959). A statistical evaluation of the Edwards Personal Preference Schedule. *Journal of Applied Psychology, 43*, 355–359. doi:10.1037/h0041451

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lucio, E., Ampudia, A., Duran, C., Leon, I., & Butcher, J. N. (2001). Comparison of the Mexican and American norms of the MMPI–2. *Journal of Clinical Psychology, 57*, 1459–1468. doi:10.1002/jclp.1109

MacDonald, A. M., & Jones, J. A. (2009). Functional imaging in clinical assessment? The rise of neurodiagnostics with MRI. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 364–395). New York, NY: Oxford University Press.

Manos, N. (1995). Adaptation of the MMPI in Greece: Translation, standardization, and cross-cultural comparison. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 4, pp. 159–207). Hillsdale, NJ: LEA Press.

McCrae, R. R., & Costa, P. T., Jr. (1983). Joint factors in self-reports and ratings: Neuroticism, extraversion, and openness to experience. *Personality and Individual Differences, 4*, 245–255. doi:10.1016/0191-8869(83)90146-0

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516. doi:10.1037/0003-066X.52.5.509

McCrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO–PI–3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment*, *84*, 261–270. doi:10.1207/s15327752jpa8403_05

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322–345. doi:10.1177/1094428106289393

Mirza, L. (1973). *Cultural adaptation, validation, and standardization of the Minnesota Multiphasic Personality Inventory (MMPI)*. Unpublished doctoral dissertation, University of Punjab, Pakistan.

Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Murray, H. A. (1938). *Explorations in personality*. New York, NY: Oxford University Press.

Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI–2 Restructured Clinical Scales. *Journal of Personality Assessment, 87*, 121–138. doi:10.1207/s15327752jpa8702_02

Nichols, D. S., Williams, C. L., & Greene, R. L. (2009, March). *Gender bias in the MMPI–2 Fake Bad Scale (FBS) and the FBS–R in the MMPI–2–RF*. Paper session presented at the meeting of the Society for Personality Assessment, Chicago, IL.

Okazaki, S., Okazaki, M., & Sue, S. (2009). Clinical personality assessment with Asian Americans. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 377–395). New York, NY: Oxford University Press.

Paterson, D. G., Williamson, E. G., & Schneidler, G. G. (1938). *Student guidance techniques: A handbook for counselors in high schools and colleges*. New York, NY: McGraw-Hill.

Piedmont, R. L., McCrae, R. R., & Costa, P. T. (1992). An assessment of the Edwards Personal Preference Schedule from the perspective of the five-factor model. *Journal of Personality Assessment, 58*, 67–78. doi:10.1207/s15327752jpa5801_6

Pope, K. S., Butcher, J. N., & Seelen, J. (2006). *The MMPI/MMPI–2/MMPI–A in court* (3rd ed.). Washington, DC: American Psychological Association.

Quevedo, K. M., & Butcher, J. N. (2005). The use of MMPI and MMPI–2 in Cuba: A historical overview from 1950 to the present. *International Journal of Clinical and Health Psychology, 5*, 335–347.

Ranson, M., Nichols, D. S., Rouse, S. V., & Harrington, J. (2009). Changing or replacing an established personality assessment standard: Issues, goals, and problems, with special reference to recent developments in the MMPI–2. In J. N. Butcher (Ed.), *Handbook of personality assessment* (pp. 112–139). New York, NY: Oxford University Press.

Reise, S. P., & Waller, N. (1993). Traitedness and the assessment of response pattern scalability. *Journal*

of Personality and Social Psychology, 65, 143–151. doi:10.1037/0022-3514.65.1.143

Rogers, R., Sewell, K. W., Harrison, K. W., & Jordan, M. J. (2006). The MMPI–2 Restructured Clinical scales: A paradigmatic shift to scale development. *Journal of Personality Assessment, 87*, 139–147. doi:10.1207/s15327752jpa8702_03

Rouse, S. V., Greene, R. L., Butcher, J. N., Nichols, D. S., & Williams, C. L. (2008). What do the MMPI–2 Restructured Clinical Scales reliably measure? Answers from multiple research settings. *Journal of Personality Assessment, 90*, 435–442. doi:10.1080/00223890802248695

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*, 168–182. doi:10.1037/0022-3514.94.1.168

Simms, L. J., Casillas, A., Clark, L. A., Watson, D., & Doebbeling, B. N. (2005). Psychometric evaluation of the Restructured Clinical Scales of MMPI–2. *Psychological Assessment, 17*, 345–358. doi:10.1037/1040-3590.17.3.345

Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 111–121). Mahwah, NJ: Erlbaum.

Sloore, H., Derksen, J., de Mey, H., & Hellenbosch, G. (1996). The Flemish/Dutch version of the MMPI–2 for Belgium and the Netherlands. In J. N. Butcher (Ed.), *Adaptations of the MMPI–2* (pp. 329–349). Minneapolis: University of Minnesota Press.

Smith, G. T., & Zapolski, T. C. B. (2009). Construct validation of personality measures. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 81–98). New York, NY: Oxford University Press.

Strack, S., & Millon, T. (2007). Contributions to the dimensional assessment of personality disorders using Millon's model and the Millon Clinical Multiaxial Inventory (MCMI–III). *Journal of Personality Assessment, 89*, 56–69. doi:10.1080/00223890701357217

Tellegen, A., & Ben-Porath, Y. S. (2008). *MMPI–2–RF technical manual.* Minneapolis: University of Minnesota Press.

Tellegen, A., Ben-Porath, Y. S., McNulty, J., Arbisi, P., Graham, J. R., & Kaemmer, B. (2003). *MMPI–2: Restructured clinical (RC) scales.* Minneapolis: University of Minnesota Press.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778–793. doi:10.1177/0013164408324460

Thorson, J. A., & Powell, F. C. (1992). Vagaries of college norms for the Edwards Personal Preference Schedule. *Psychological Reports, 70*, 943–946. doi:10.2466/pr0.1992.70.3.943

Wallace, A., & Liljequist, L. (2005). A comparison of the correlational structures and elevation patterns of the MMPI–2 Restructured Clinical (RC) and Clinical Scales. *Assessment, 12*, 290–294. doi:10.1177/1073191105276250

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132–1139. doi:10.1111/j.1467-9280.2009.02417.x

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology* (Vol. 10, pp. 43–66). New York, NY: Wiley. doi:10.1002/0471264385.wei1003

Wiggins, J. S. (1969). Content dimensions in the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 127–180). New York, NY: McGraw-Hill.

Williams, C. L., Butcher, J. N., Gass, C. S., Cumella, E., & Kally, Z. (2009). Inaccuracies about the MMPI–2 Fake Bad Scale in the reply by Ben-Porath, Greve, Bianchini, and Kaufman (2009). *Psychological Injury and Law, 2*, 182–197. doi:10.1007/s12207-009-9046-3

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Woodworth, R. S. (1920). *Personal data sheet.* Chicago, IL: Stoelting.

Worell, J., & Robinson, D. A. (2009). Issues in clinical assessment with women. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 415–431). New York, NY: Oxford University Press.

Zapata-Sola, A., Kreuch, T., Landers, R. N., Hoyt, T., & Butcher, J. N. (2009). Clinical personality assessment in personnel settings using the MMPI–2: A cross-cultural comparison. *International Journal of Clinical and Health Psychology, 9*, 287–298.

CHAPTER 12

CLINICAL ASSESSMENT: A MULTICULTURAL PERSPECTIVE

Lisa A. Suzuki, Mineko Anne Onoue, and Jill S. Hill

Clinical assessment is part of the core of professional practice in psychology. The sheer number of tests in a variety of psychological domains has increased exponentially over many decades. *The Buros Institute of Mental Measurements* provides reviews of over 11,000 tests (or test revisions) published since 1935, with the majority of these measures originating in Western societies such as the United States, Western Europe, Canada, and Australia (Oakland, 2009).

As the practice of testing has grown, so have the number of concerns regarding the application of personality and intelligence measures with individuals from diverse cultural backgrounds. In particular, charges of cultural bias and questions regarding equivalence, reliability, and validity have been raised as usage of the most popular measures have spread to new populations for which the tests were not originally designed.

Despite these concerns, psychological assessment remains a mainstay for many clinicians working with members of diverse cultural groups. Many test developers make concerted efforts to create measures that are applicable to different racial and ethnic groups. These efforts include using explicit procedures to determine content validity, addressing potential bias through oversampling of particular groups, and using statistical procedures to address potential concerns. In addition, enhanced computer technology has enabled rapid shifts in the assessment process through implementation of alternative testing formats (e.g., item response theory) and methods to discern equitable assessment across diverse populations, taking into consideration testing procedures, scoring, and use of scores (Dana, 2000; Mpofu & Ortiz, 2009).

For decades, scholars have noted the importance of understanding cultural context as it affects psychological assessment from start to finish. Cultural competence requires that the examiner possess cultural awareness, knowledge, and skills (American Psychological Association, 2002). This chapter focuses on the following areas: (a) cultural challenges in clinical assessment, (b) cultural competencies in assessment, (c) integrating culture in the process of assessment, and (d) cultural considerations in the use of popular personality and intelligence measures.

CULTURAL CHALLENGES IN CLINICAL ASSESSMENT

One of the greatest challenges facing clinicians has been the integration of cultural context into the process of assessment. The use of psychological tests with diverse populations in the United States (e.g., immigrants and refugees), as well as the exporting of tests to new cultural contexts (e.g., countries), increases the need for culturally appropriate assessment practices (see Volume 3, Chapters 11 and 26, this handbook, for more information on test use with children across cultures). The complexity of the clinical assessment process is amplified by the multiple and subtle nuances of culture. Indeed, culture

We thank Muninder K. Ahluwalia and John F. Kugler for their feedback on earlier versions of this chapter.

DOI: 10.1037/14048-012
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

encompasses a number of unique identities that are not limited to the sole emphasis often placed on race or ethnicity. These aspects include experiences related to geographic boundaries, language, religious belief, social class, gender, sexual orientation, age, and ability status (Goldberger & Veroff, 1995). Cultures and their related identities are dynamic and changing (e.g., López & Guarnaccia, 2000), further complicating our understanding of the individual in the assessment process. Although assessment professionals understand the limitations of examining the testing literature focusing on race and ethnicity as indicative of culture, this practice has some value and remains at the forefront of the literature in the area of psychological assessment.

MULTIPLE DEFINITIONS: A SOURCE OF CULTURAL CONFUSION

More than a decade ago, Pedersen (1999) noted that culture affects the measurement of all psychological phenomena, yet it remains one of the "most misunderstood constructs in contemporary theories of psychology" (p. 3). Kroeber and Kluckhohn (1963) noted more than 100 meanings in their seminal review of the literature on "culture." According to the American Psychological Association (APA; 2002), *culture* includes reference to belief systems, value orientations (i.e., customs, norms, practices, and social institutions), psychological processes (e.g., language, caretaking practices, media, educational systems, organizations), worldview, learned and transmitted beliefs, and practices (e.g., religious and spiritual traditions). Widely accepted definitions such as this one are general and difficult to operationalize. All individuals are considered to be "cultural beings," emphasizing racial and ethnic heritage. Given the breadth of this definition, it is no wonder that difficulties arise in studying this important construct in psychological research and that there is continued reliance on the seemingly simpler racial and ethnic group designation.

Similar problems have arisen with respect to the most popular measurement constructs. For nearly a century, standardized intelligence testing has been identified as one of psychology's "greatest successes" and "most persistent and widely used inventions" (Benson, 2003, p. 48). Although several theories of intelligence have emerged in the literature, some of the most widely used measures in the area lack a clear theoretical foundation. Thus, Boring's (1923) conclusion that intelligence is what intelligence tests measure is still relevant today. Difficulties in operationalizing the definition of culture complicate the assessment process. As noted earlier, relying on self-identified categories such as one's racial and ethnic group membership does not capture the full complexity of multiple cultural identities. This shortcoming, coupled with the atheoretical development of particular psychological measures, creates complex challenges for the clinician conducting a psychological assessment.

RAPID EXPANSION OF MEASURES AND TESTING PRACTICES

In addition to challenges in construct definition, problems have arisen due to rapid acceleration in the development of psychological measures in a variety of domains aimed at meeting societal demands for quick and efficient classification of individuals and accountability. One of the origins of the psychometric movement in psychology was spurred by historical pressures to create tests to classify individuals during wartime (see Chapter 8, this volume, which provides a historical overview of the development of measures of intellectual ability). The Army Alpha and Beta tests were the precursors to the first intelligence tests. Currently, the most frequently used instruments in the United States are transported globally and often renormed and restandardized on new populations creating new challenges. Accordingly, Cheung and Cheung (2003) observed that personality assessment is an "import" in Asian countries.

Tests have become one of the major gatekeepers to higher education opportunities. High-stakes testing is now a central component in the lives of students in the United States. Test preparation has an increased presence in the testing industry. Companies have capitalized on this movement and, for a price, provide students and others with training in test-taking strategies, in some instances, guaranteeing improved performance. Test-taking strategies are also taught as part of the curriculum in many

schools. In addition, parents routinely invest in private tutoring to prepare their children for these high-stakes tests.

Test scores are also used to compare the quality of education between various countries. Results from the Program for International Student Assessment (PISA; Organisation for Economic Co-operation and Development [OECD], 2009) were released in 2009 ranking the 65 participating industrialized countries based on scores in math, reading, and science. Topping the list was Shanghai, People's Republic of China, prompting educational leaders in the United States to compare the culture of education between the different countries (Dillon, 2010). Reports, however, also indicated that the students in Shanghai are not a representative sample of China overall. Despite the sampling limitation, the U.S. Secretary of Education, Arne Duncan, commented, "We have to see this as a wake-up call. . . . The United States came in 23rd or 24th in most subjects. We can quibble, or we can face the brutal truth that we're being outeducated" (cited in Dillon, 2010).

There is little doubt that testing in the United States has a major role in educational and psychological practices. However, the effect of these measures in various cultural contexts here as well as abroad remains controversial. Understanding the strengths and limitations of our assessment tools and practices is imperative to ensure that results are accurate and useful to the lives of individuals from diverse cultural backgrounds.

TEST BIAS

All psychological measures are developed within particular cultural contexts and problems have arisen when measures are used with different cultural groups and assumed to have multicultural or universal application. Historically, assumptions of universality (i.e., the assumption of cross-cultural validity or "imposed etic strategy"; Berry, 1989) have been controversial given charges of test bias. Test bias specifically pertaining to race or ethnicity is often known as *cultural bias* (Valencia, Suzuki, & Salinas, 2000). Racial and ethnic group differences in scores obtained on major personality and intelligence tests were identified and led to criticisms in the usage of these measures. For example, differences in subtest profiles on the Minnesota Multiphasic Personality Inventory (MMPI) and Wechsler scales were used as indicators of cultural bias.

Reynolds (1982, as cited in Reynolds & Lowe, 2009) stated that bias is a statistical term referring to "systematic error in the measurement of a psychological attribute as a function of membership in one or another cultural or racial subgroup" (p. 333; see also Volume 1, Chapter 8, this handbook, for more information on bias in testing). Sources of bias may include those related to the test itself (e.g., inappropriate content and standardization samples, differential predictive validity), those introduced during the usage of a particular measure (e.g., examiner or clinician, language), and others based on how test results are used (e.g., determination of service delivery, inequitable social consequences; Dana, 2005; Reynolds & Lowe, 2009). There also exist different forms of bias including outcome bias, predictive bias, and sampling bias (Serpell, 2000). Studies of cultural bias in testing have yielded mixed results depending upon the construct being assessed and the test being examined (e.g., Valencia & Suzuki, 2001).

CULTURAL COMPETENCIES IN ASSESSMENT

The psychological assessment of members of diverse communities and the growing awareness of the importance of cultural context has led to the creation of a number of professional guidelines and delineation of multicultural competencies related to clinical practice. These include the APA's *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010; see also Volume 1, Chapter 15, this handbook, which highlights ethics in psychological assessment) and the *Guidelines on Multicultural Education, Training, Research, Practice and Organizational Change for Psychologists* (APA, 2002) as well as assessment competencies delineated by Krishnamurthy et al. (2004). Attention to cultural differences is also noted throughout the ongoing revision of the *Standards for Educational and Psychological Testing*, endorsed jointly by the American Educational Research Association, APA, and the National Council on Measurement and Evaluation (2012), that is

currently under way. The assessment work group of the 2002 Competencies Conference: Future Directions in Education and Credentialing in Professional Psychology included attention to the complexities of a multicultural model of psychological assessment with an emphasis on training clinicians (Krishnamurthy et al., 2004).

All of these publications and policy statements highlight the importance of understanding the findings of psychological assessment within a cultural context. From start to finish, the assessment process requires that the clinician be knowledgeable regarding the cultural background of the individual and its potential effect on test performance, aware of his or her own unique cultural identity and how perceptions of his or her role may affect the testing relationship, and skillful in conducting and interpreting the results of the assessment process.

INTEGRATING CULTURE IN THE PROCESS OF ASSESSMENT

Assessment should be tailored to the unique characteristics of the individual being assessed. The following sections highlight traditional steps in the assessment process highlighting cultural considerations (see Chapter 2, this volume, for more information on the general assessment process).

DETERMINING THE REASON FOR REFERRAL

Determining the reason for referral can be a complex process requiring input from various sources including the individual being assessed and the referring agent. Presenting problems can involve a number of issues pertaining to the individual as well as the community in which they reside. It is critical that problems be understood within the cultural context and life circumstance of the person being assessed (Flores & Obasi, 2003).

OBTAINING INFORMED CONSENT

Informed consent is a cornerstone of Western social science. However, attention to issues related to multicultural and multiethnic populations remains limited (Guerrero & Heller, 2003) with respect to this process, which involves the sharing of information between the clinician and the client (Hoop, DiPasquale, Hernandez, & Roberts, 2008). Cultural background may influence the ability of an individual to voluntarily give consent. One issue can be an individual's level of language proficiency. For example, consent forms often contain "jargon" that may be unfamiliar to the participant (Waggoner & Mayo, 1995). In addition, clinicians must consider the reading level of the consent form itself. Language, in conjunction with educational level, literacy, and socioeconomic status, has been identified as a barrier in the consenting process (Guerrero & Heller, 2003). Scholars have also noted that "rigid adherence to North American norms for informed consent can violate both subjects' and researchers' culture-specific communication codes in societies where human relations function differently from U.S. or European habitual patterns" (Hong, 1998, p. 81). The act of requiring participants to place their signature on contractual documents using phases such as "I understand," "I am aware of," and "I have a right to" may not have relevance for the client and can raise additional questions rather than alleviate potential concerns (Hong, 1998). Thus, the evaluator must ensure the cultural validity of the informed consent process so that the examinees' rights are protected on the basis of their sociocultural reality (Hong, 1998).

It is also important to recognize the power that the clinician wields that may lead the client to agree to whatever the clinician requests. It must be noted that tacit approval is not acceptable; consent must be affirmative and explicit. Making the process of assessment as transparent as possible is critical to establishing open communication with clients. Greater transparency increases the likelihood that they will be active participants in the assessment process.

CULTURE AND THE CLINICAL INTERVIEW

Obtaining relevant information regarding the individual examinee's history, social location, and cultural background is imperative in the assessment process and can most often be captured through a

qualitative interview (see Chapter 7, this volume, for more information on the clinical interview).

A number of interview protocols have been developed to assist the psychologist in gaining important information regarding the cultural background of the individual being evaluated. Two such interviews are the Person-in-Culture Interview (PICI; Berg-Cross & Takushi-Chinen, 1995) and the Cultural Assessment Interview Protocol (CAIP; Grieger, 2008). The PICI was developed to assist professionals in developing cross-cultural understanding. This 24-item interview requires participants to share their worldview. The interview attends to both cultural and idiosyncratic values from psychodynamic, humanistic, family, and existential perspectives. Questions address pleasurable activities, the effect of the problem on the self and family members, familial expressions of emotion, finances, safety, roles, experience in the community, important people in the individual's life, religious beliefs, meaning attached to experiences, and responsibilities.

The CAIP (Grieger, 2008) is based on the premise that examining cultural issues in the process of assessment is appropriate for all clients. In Part I, Gathering Cultural Data, the CAIP includes questions regarding problem conceptualization, cultural identity, level of acculturation, family structure and expectations, level of racial/cultural identity development, experiences with bias, immigration issues, existential/spiritual issues, and counselor characteristics and behaviors (e.g., aspects of the counselor's identity that are salient to the client, client's perception of positive and negative behaviors of the counselor). In Part II, Integrating Cultural Data, the clinician addresses the implications of cultural factors that exist in the relationship between the counselor and the client and puts forth a summary of cultural factors and implications for diagnosis, case conceptualization, and treatment.

Quantitative instruments are also available to assist the clinician in obtaining information about salient constructs highlighted in the clinical interview, with emphases on acculturation, socioeconomic status, and language. Many of these instruments are prominently featured in the literature but have not become part of a standardized assessment protocol. It is important, however, to consider these constructs when conducting a psychological evaluation.

Acculturation

Acculturation refers to "a dynamic process of change and adaptation that individuals undergo as a result of contact with members of different cultures. This change is influenced by the environment the individual resides in, as well as, qualities innate to that individual" (Rivera, 2008, p. 76). Acculturation has been linked to a number of psychological issues including mental health, academics, and family relationships and functioning (Rivera, 2008; see also Chapter 23, this volume, for more information on acculturation).

Kim and Abreu (2001) noted that the definition of acculturation has evolved over the years from one that was initially conceptualized as a unilinear process (i.e., minority individuals adapting to the mainstream culture) to one that is multilinear, encompassing multiple settings and cultures. Berry (2003) noted the existence of four acculturative strategies: assimilation, separation, marginalization, and integration.

A functional definition of acculturation encompasses behavior, values, knowledge, and affective cultural identity (Kim & Abreu, 2001). The behavioral level of functioning includes friendship choice, TV and reading preferences, participation in cultural activities, and contact with indigenous culture. The cognitive level includes values (i.e., attitudes and beliefs about social relationships, cultural customs and traditions, gender roles, attitudes about health and illness) and knowledge (i.e., culturally and historically specific information related to culture of origin and dominant culture, and meaning attached to culturally specific activities). The affective level includes cultural identity (i.e., attitudes, cultural identification, and attitudes toward indigenous and dominant groups).

Socioeconomic Status (SES)

SES refers to the social standing or class of an individual or group. It is often measured as a combination of education, income, occupation, access to resources, privilege, power, and control (APA, n.d.).

SES is a multidimensional construct that has been positively linked to the measurement of health, psychological well-being, and "attainment of social and culturally derived goals" (Ensminger & Fothergill, 2003, p.13). The three most common indicators of SES are parental educational achievement (especially maternal), financial income, and occupational status. In addition to these individualized indicators of SES, there is growing attention to characteristics of households and neighborhoods that influence the availability of resources and living conditions.

Concerns have arisen given the mounting evidence that socieconomic indicators may not have the same meaning for immigrant families as they do for U.S.-born families. For instance, Fuligni and Yoshikawa (2003) have emphasized the importance of contextualizing SES in terms of the history of U.S. immigration policies (i.e., entry preferences and quotas, provisions for refugees and asylum seekers, and eligibility for federal assistance) for immigrant families.

Immigrants from Asian countries tend to possess higher levels of education, work in higher status jobs, and have significantly more income than those from Latin America. Those from Africa are approximately equal to their Asian counterparts in terms of education, but they tend to work in lower status occupations and earn lower incomes. The low SES of Latin Americans is largely because only one third of those from Mexico have graduated from high school (Fuligni & Yoshikawa, 2003).

Fuligni and Yoshikawa have further noted the importance of understanding the socioeconomic resources of immigrant families in terms of human capital (e.g., nonmaterial resources, including cognitive stimulation, and values pertaining to achievement, such as parental educational level), financial capital (e.g., physical resources including wealth and income), and social capital (e.g., resources available through relationships and connections in family and community) given the limitations of traditional measures, concluding that, "traditional indicators of human and financial capital can be problematic for immigrant families because these indicators may simultaneously underestimate and overestimate the resources available to parents and children" (Fuligni & Yoshikawa, 2003, p. 111). For example, parents' educational levels may vary given that educational opportunities may have been limited in their original homeland.

Language

Most instruments in the United States are developed for native English-language speakers. The complexities of assessment considering language differences were noted by Ortiz and Dynda (2005):

> Individuals who are bilingual either by circumstance or choice, are significantly different and do not have background experiences that are comparable to the monolingual individuals who comprise existing norm samples. Bilinguals are different not only from monolingual English speakers but also from monolingual native language speakers so that tests that utilize one group or other for comparison purposes remain equally inadequate. (p. 554)

The Multidimensional Assessment Model for Bilingual Individuals (Ortiz & Ochoa, 2005) addresses the complex features of linguistically diverse individuals. The process involves the collection of data from multiple sources highlighting cultural and linguistic history in the areas of language, instructional programming, and current grade level. These data are used to assist the clinician in determining the most appropriate assessment method—nonverbal assessment, assessment in native language, assessment in English, or bilingual assessment (see also Volume 3, Chapter 17, this handbook, for more information on the assessment of English-language learners).

In addition, clinicians may use various web-based tools to help them in determining the readability level of a particular measure under consideration. For example, the website for Intervention Central (2011) contains a curriculum-based measure maze passage generator that calculates the reading level of passages consisting of 75 words or more utilizing 10 different formulas to calculate the grade level of the passage. The examiner chooses the formula closest to their clients' characteristics (see Volume 3, Chapter 8, this handbook, for more information on curricular assessment).

CULTURE IN BEHAVIORAL OBSERVATIONS

Behavioral observations provide another important source of information in clinical assessment. These observations, whether in the form of narrative records or behavior checklists, can provide an additional dimension to verbal reports (Sattler & Hoge, 2006). Behaviors must be understood within the cultural context of the individual, as the meaning of overt actions may be culturally influenced. Sattler and Hoge (2006) emphasized the importance of "sensitivity, acuity and perceptiveness" (p. 193) on the part of the observer. Behavioral expressions of emotion and aspects of nonverbal communication may be influenced by cultural norms (Sue & Sue, 2008). For example, Sue and Sue (2008) noted that Japanese children may appear quiet and reserved because they have been taught not to speak until someone speaks to them. This behavioral tendency may lead a clinician to conclude incorrectly that a Japanese child is inarticulate and unintelligent. Understanding the overt and sometimes subtle ways that culture affects behavior is important in gaining a genuine understanding of an individual's behavior in proper context.

SELECTING APPROPRIATE MEASURES

If an assessment is deemed necessary, the evaluator must select appropriate measures to address the identified problem(s) or reason(s) for referral. Flores and Obasi (2003) noted that, in selecting instruments, it is critical to examine whether the construct being measured is present in the individual's cultural context and whether it has the same meaning. These questions highlight the importance of conceptual and functional equivalence. These authors further maintained that "because of cultural variance in behaviors, customs, and norms, the selection of measures should consider not only the definition of the construct but also how the construct would be manifested in an individual's culture" (Flores & Obasi, 2003, p. 45). As noted in the earlier section on language, if the individual is not proficient in English, then the examiner must review measures available in the examinee's language or measures without language requirements (i.e., nonverbal tests). Translated versions of frequently used measures also may be considered, if available.

CULTURAL ADAPTATION AND TRANSLATION OF TESTS

Hambleton (2005) noted that the process of test adaptation enables the clinician to utilize a measure created in one cultural context and transport it to another:

> Test adaptation includes all the activities from deciding whether a test could measure the same construct in a different language and culture, to selecting translators, to deciding on appropriate accommodations to be made in preparing a test for use in a second language, to adapting the test and checking its equivalence in the adaptive form. (p. 4)

Translation is part of the adaptation process and involves much more than merely attending to the literal written translations of item content (in contrast to interpreters, who focus on oral language). For example, "translators are trying to find concepts, words and expressions that are culturally, psychologically, and linguistically equivalent in a second language and culture" (Hambleton, 2005, p. 4). Caution should be used in administering translated versions, given that they may not be renormed or restandardized on a population relevant to the individual being assessed. Lopez (2010) noted that translating measures involves a number of important procedures to address issues of validity beyond just translating the test items. She noted that an editorial review committee should examine the translated version and that the draft measure should be adapted based on their feedback. The measure should then be pilot tested and field tested before standardizing the scores and performing validity studies. More information regarding test adaptation is available in the International Test Commission Guidelines for Test Adaptation (cited in Hambleton, 2005) and in Volume 3, Chapter 26, this handbook.

INTEGRATING ALTERNATIVE METHODS OF ASSESSMENT

In the process of data gathering, it is important for the examiner to consider and integrate, when appropriate, alternative forms of assessment. This action may include measures with a variety of response formats. For example, some tests may allow a written response or an oral response; others may select from multiple choices or forced choices (e.g., recognition), drawing, and so forth.

In addition to alternative response formats, examiners should be aware of the multifactorial nature of most measures (i.e., tasks often involve multiple skill areas) that may make it difficult to ascertain explanations for performance on particular tasks. Therefore, the evaluator must consider task demands to determine what will yield the most informative data related to the referral. Ochoa (2003) highlighted literature regarding alternative assessment methods for linguistically diverse students. These practices include direct observation, rating scales, checklists, performance assessments, work samples, student interviews, criterion-referenced tests, curriculum-based measurement, and dynamic assessment.

Administering the Assessment

The relationship between the examiner and the examinee in an assessment is a fleeting one and is often established solely for the purposes of the evaluation. Nevertheless, it is critical that the examiner establish rapport with the individual being tested and other stakeholders. Sattler and Hoge (2006) noted that the evaluator should communicate his or her knowledge of the individual's cultural background and allow time to establish a trusting relationship. Knowledge of cultural norms and mores should be used to inform test administration. For example, McShane (1980) has supported sociolinguistic modifications in the assessment of American Indian children. Specifically, examiners should avoid looking directly at children; accommodate lower levels of speech; and be aware of the tendency of children to use short, quick responses. He has also suggested that examiners sit across from, but not right in front of, the child being tested so that they can observe the child without having to stare directly at them.

Potential problems can arise in communication between the clinician and the examinee that may affect the validity of the results (Hambleton, 2005). Instructions should be clearly presented without relying solely on verbal language. Hambleton (2005) has cited research cautioning the usage of rating scales and multiple-choice formats that may be unfamiliar in some cultural contexts. In addition, some measures rely on the examinee's ability to complete tasks quickly and efficiently. Working quickly "may not be known or understood by examinees in different cultures" (Hambleton, 2005, p. 9).

Testing the Limits

Although standardized administration protocols always should be followed to ensure the validity of results, psychologists may opt to use follow-up procedures to gain additional information. Testing the limits can include readministering items with additional supports (e.g., providing paper and pencil for orally presented word problems).

Sattler and Hoge (2006) noted the following in their discussion of procedures in the assessment of children with brain injuries:

> Testing-of-limits may include modifying instructions to involve more or fewer cues, adjusting the pace at which information is presented, modifying the modality of presentation, modifying the starting or discontinuance procedures by administering additional items, adjusting memory demands (e.g., using recognition instead of recall procedures), modifying the response format (e.g., allowing pointing instead of oral responses), adjusting task complexity (e.g., making tasks more concrete), and asking for explanations of responses. (p. 557)

After the test is administered according to its standardization, the examiner may allow additional time for completion of a task, ask the examinee for an oral explanation of a particular response, or provide paper and pencil in cases where standardization requires that the individual solve problems without these aids. It should be noted that any deviation

from the standardization protocol should be duly noted in the written report of the assessment results.

Testing-the-limit procedures can be helpful in obtaining information about how a student may perceive and approach the tasks involved in an assessment. For example, think-aloud protocols (TAPs; Ercikan et al., 2010) can be used to identify cognitive differences that may exist between test takers from different backgrounds. TAPs are conducted to better understand examinees' thinking processes. This technique involves having test takers work on problems that are presented in the assessment while they verbalize their thinking processes either as they solve the problem or after they have finished answering the question. To understand accurately the cognitive processes that are taking place, examiners may also provide verbal prompts when necessary. Dynamic assessment procedures that follow a test–teach–test format may also be a helpful approach. Students are presented with a task to gain a baseline measure of performance, then taught by the examiner how to approach and solve the problems. The student is then tested again to gain a closer approximation of their true ability (see Volume 3, Chapter 7, this handbook, which elaborates on dynamic assessment procedures).

Usage of Interpreters

As noted earlier, when psychologists do not have bilingual and multilingual skills, they must often rely on interpreters. In addition to being fluent in two or more languages, these individuals must be trained to understand the assessment process, the technical language used in the items, and ethical issues that may affect the evaluation (Lopez, 2010). Information is available regarding standards of practice and ethics put forth by the National Council on Interpreting in Health Care (2004, 2005). Each of these publications includes a specific emphasis on the importance of cultural understanding, language proficiency, and clarification in instances of cultural misunderstanding.

INTEGRATING CULTURE IN INTERPRETATION OF TEST RESULTS

Interpreting results requires the ability to analyze and synthesize information from an array of data sources taking into consideration the cultural context of the individual. These include the interview(s), observations, anecdotal records, and results obtained from the tests themselves.

Communicating the results of a psychological evaluation in a formal report and in person to the client is often seen as the final step in psychological testing. It is important that information is presented in a jargon-free manner. Focusing on the individual's strengths as well as areas of limitation is an important part of the process. Finding areas in which an individual works well may be an important key to strategizing future learning and intervention strategies. In addition, allowing the client to comment on the interpretation is essential and underscores the need to maintain rapport throughout the entire assessment process. Readers are referred to Chapter 3 in this volume for further discussion concerning the communication of test findings.

FORMULATING RECOMMENDATIONS

The evaluator must be astute at problem solving and knowledgeable about the particular setting(s) and interventions available to the individual within their community. Recommendations must be informed and realistic. They must include attention to the individual's social location and community context. For example, recommendations for a child having learning difficulties must take into consideration cultural factors as well as state-of-the-art and cutting-edge interventions. Psychologists also should have knowledge regarding the resources needed and where these services can be accessed. At times, the examiner must note the best available solution, given difficulties with community access and limited resources. In such cases, the role of the examiner as advocate may become paramount.

CULTURAL CONSIDERATIONS IN THE USAGE OF PERSONALITY MEASURES

There are different conceptions of what constitutes the relationship between culture and personality. Piekkola (2011) has offered the following: "Rather than conceiving of personality as fixed and universal, it is argued that personality is an adaptation

worked out in the cultural and historical context of the individual life" (p. 1). The Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), Millon Clinical Multiaxial Inventory—III (MCMI–III; Millon, Millon, Davis, & Grossman, 2006), Thematic Apperception Test (TAT; Murray, 1943), and the Rorschach (Exner, 2003), are among the most frequently used measures of personality currently in existence in the United States (Camara, Nathan, & Puente, 2000). Each of these measures will be discussed in turn, with particular attention to cultural issues in their use with diverse clientele (see Chapter 24, this volume).

The MMPI–2

The MMPI–2 is the most widely used and researched personality assessment instrument in the world (Butcher & Ben-Porath, 2004; Dana, 2000, as cited in Hill, Pace, & Robbins, 2010). This esteemed status is largely due to the extent to which it has been empirically validated (Butcher, 2000; readers also may consult Chapter 11, this volume). Researchers who have attempted to determine whether the MMPI–2 is biased have pointed to the identification of empirical correlates as the best method to address the issue (Arbisi, Ben-Porath, & McNulty, 2002; Greene, Robin, Allbaugh, Caldwell, & Goldman, 2003; Timbrook & Graham, 1994). However, this stance does not take into account that extratest measures used in the identification process are typically rooted in the same epistemological framework or worldview as the MMPI–2 (Pace et al., 2006). In fact, this stance uncritically assumes that such measures are universally applicable, leading to a tautological process whereby measures developed from the dominant cultural worldview are utilized to empirically validate the MMPI–2, also developed from that dominant cultural worldview (Hill, Robbins, & Pace, 2012). Therefore, it appears that the findings from research that approaches the issue of test bias in this manner are suspect.

A better practice is the use of extratest measures that are grounded in the nondominant group's culture to empirically validate the MMPI–2. This type of effort lends cultural credibility and validity not only to research findings but also, more important, to the research process and methods used (Hill et al., 2012).

Groth-Marnat (2009) summarized the research regarding use of the MMPI–2 with diverse groups. He noted that several reasons exist for differences in scoring patterns between racial and ethnic groups:

> Although scores may be due to the accurate measurement of different personality traits, they may also be the result of cultural tendencies to acquiesce by giving socially desirable responses, different beliefs about modesty, role conflicts, or varying interpretations of the meaning of items. Profiles may also reflect the results of racial discrimination in that scales associated with anger, impulsiveness, and frustration may be elevated. (p. 221)

A number of studies have been conducted that examine racial and ethnic group differences on various editions of the MMPI. Seminal articles have indicated significant differences between African Americans and Caucasians across scales (Castro, Gordon, Brown, Anestis, & Joiner, 2008; Greene, 1987; Hall, Bansal, & Lopez, 1999). In his comprehensive review, Groth-Marnat (2009) noted that African Americans are more likely to score higher on Scales F, 8, and 9. He noted the importance of understanding "moderator variables, such as education, income, age, and type of pathology" (p. 221). In addition, when discrepancies have been found, the point difference is often not clinically meaningful (e.g., Hall et al., 1999).

Native Americans as a group score higher on Scales L, F, K, 4, 8, and 9 (Robin, Greene, Albaugh, Caldwell, & Goldman, 2003). Research suggests that, when the score differences between native samples and the norm group are small and below a *T* score of 65, it is more likely that the differences are due to cultural differences and experiences with oppression rather than psychopathology. Score elevations above 65 may be indicative of psychopathology (Groth-Marnat, 2009; Hill et al., 2010; Pace et al., 2006; Robin et al., 2003).

Several research studies conducted on Asian groups focused on test development and application of the MMPI–2 internationally in Asian countries

(Butcher, Cheung, & Kim, 2003; Cheung, Zhao, & Wu, 1992; Hahn, 2005; Ketterer, Han, Hur, & Moon, 2010; Sukigara, 1996). Other studies targeted the effects of acculturation on MMPI–2 scores (Tsai & Pike, 2000). Less acculturated samples in the United States have been found to score higher on a number of scales in relation to those assessed to be more acculturated (Sue, Keefe, Enomoto, Durvasula, & Chao, 1996; Tsai & Pike, 2000).

Studies focusing on Latino groups are limited, given a lack of consistent findings. In a meta-analysis of 19 studies (Hall et al., 1999), findings based on aggregate effect sizes indicated that Latino American males scored higher on three validity scales (L, F, and K) and obtained lower scores on all clinical scales (1, 2, 3, 4, 5, 6, 7, 8, 9, and 0), although the effect sizes were small. The effect sizes were "robust only for Scales L and 5" (Hall et al., 1999, p. 191), indicating that Latinos scored higher on the L scale and lower on Scale 5, although total score point differences were less than 5 *T*-score points. Studies addressing the Latino/a population generally recommend MMPI–2 usage with caution.

The results of these studies demonstrate the possibility of what Dana (1993) termed a culture-psychopathology confound across several of the same MMPI–2 clinical and validity scales for various racial, ethnic, and cultural minority groups (Hill et al., 2010). This confound calls into question the test's validity when used with African Americans, Latinos/as, Asian Americans, and American Indians. As a result, Dana and several other researchers strongly suggest that clinicians utilize accompanying supplemental measures of racial identity or acculturation status whenever they administer the MMPI–2 in a multicultural context (Allen & Dana, 2004; Dana, 1993; Hill et al., 2010; Whatley, Allen, & Dana, 2003). The use of such measures has the possibility of illuminating the meaning behind observed score differences for various racial, ethnic, and cultural groups rather than simply reaffirming the existence of these normative differences devoid of any culturally informed context.

In support of these recommendations, Groth-Marnat (2009) suggests that, in relation to the MMPI–2, future research should address ethnic group differences in relation to "acculturation, language fluency, perceived minority status and degree to which they feel discrimination" (p. 222). Relationships between the MMPI and the status of African American racial identity have been identified in relation to Scales F, 4, 6, 8, and 9 (Whatley et al., 2003). Resources are available addressing competent practice in the use of the MMPI–2 with Latino/a populations (Velasquez et al., 1997). Finally, there appears to be greater support for the consistent use of local norms and acculturation status norms when applying the MMPI–2 to unique cultural contexts (Allen & Dana, 2004; Hill et al., 2010; Pace et al., 2006).

The MCMI–III

The MCMI–III (Millon et al., 2006) was developed to assess both enduring personality features and clinical syndromes of psychiatric and emotionally disturbed populations fusing diagnostic categories of the text revision of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000). The MCMI–III is a popular measure, one of the most frequently used in clinical practice (Archer, Buffington-Vollum, Stredny, & Handel, 2006) and taught in clinical academic settings (Piotrowski & Zalewski, 1993). The MCMI–III has demonstrated reliability and validity in research settings; however, one concern was that many ethnic groups, with the exception of American Indians, were underrepresented in the normative sample compared with U.S. Census norms (Kwan & Maestas, 2008).

Research on the use of the MCMI–III with diverse racial and ethnic group samples indicates that African American psychiatric inpatient populations were found to have higher scores on Histrionic, Narcissistic, Paranoid, Drug Dependent/Drug Abuse, and Delusional Disorder/Psychotic Delusion scales in comparison with Whites (Munley, Vacha-Haase, Busby, & Paul, 1998). When matched for primary Axis I discharge diagnosis and substance abuse comorbidity, these scale differences were reduced. In other studies, African American inpatient psychiatric patients scored higher than White patients on the Antisocial, Avoidant, Psychotic Thinking, and Psychotic Delusion scales (Davis, Greenblatt & Pochyly, 1990, as cited in Munley et al., 1998). In populations of outpatient psychiatric

court referred spouse abusers, African Americans scored higher than Whites on the Narcissistic, Paranoid, Hypomanic, Drug, Abuse, and Psychotic Delusion scales (Hamberger & Hastings, 1992, as cited in Munley et al., 1998).

More recently, Groth-Marnat (2009) reported that differences between Caucasian American and African American psychiatric patients "have been found on 9 of the 20 MCMI–III scales" (p. 306). In particular, he cited the work of Choca, Shanley, Peterson and Van Denberg (1990) indicating higher scores for African Americans on the Antisocial, Narcissistic, Paranoid, Hypomania, and Drug Abuse scales. The meaning of these score differences is unclear, although he noted that, "Accuracy of MCMI–III elevations is supported in that self-descriptions by African American clients closely correspond to expected elevations on the MCMI–III" (Craig & Olson, 2001, as cited in Groth-Marnat, 2009, p. 307). Kwan and Maestas noted that, on the basis of their review of several research studies on the racial comparison of African Americans to Caucasians, there is a lack of attention and depth given to the reasons why assessment differences based on race may or may not exist.

Reviews of research on the use of the MCMI–III with ethnic populations other than African Americans are scarce. In a comparison of Native American and non-Native American incarcerated men, Native American men were found to score higher on Compulsive, Avoidant, Schizoid, Alcohol Abuse, Thought Disorder, and Debasement scales than non-Native American men (Glass, Bieber, & Tkachuk, 1996). Kwan and Maestas (2008) questioned the appropriateness of the MCMI–III to *predict* psychological disturbance for ethnic groups other than Caucasian. Some research shows that the MCMI–III diagnoses of personality disorder were given at a much higher rate than diagnoses based on clinical interviews (Wetzler, 1990). Kwan and Maestas recommended thoughtful analysis of an individual's ethnic and racial group membership before using the MCMI–III to inform diagnoses.

The TAT

The TAT (Murray, 1943) is based on the assumption that ambiguous, emotionally provocative pictures elicit projected personal narrative that express personality structure (Gray-Little, 2009). Results from the TAT can contribute to understanding areas of emotional, interpersonal, and motivational characteristics as well as defensive mechanisms and problem-solving style (Groth-Marnat, 2009). The TAT is composed of drawings of scenes that reflect White individuals in various situations; thus, the drawings are bound in cultural norms and histories of a White European-American group (Groth-Marnat, 2009). Research indicates that the TAT is most often interpreted by the clinician in terms of a subjective impression (Lilienfeld, Wood, & Garb, 2000).

The content of the TAT has a long history of being questioned for cross-cultural use, starting in the 1950s. It has been adapted several times to better assess various racial and ethnic groups (Gray-Little, 2009). Empirically, the TAT's cross-cultural application has been questioned in terms of reliability and validity for any one particular racial or ethnic group based on the relationship between TAT results and other self-report measures (Lilienfeld et al., 2000). There is a dearth of research regarding the TAT's validity with particular ethnic and cultural groups such as Native American communities (Monopoli & Alworth, 2000), which may reflect a general consensus in the field that the TAT in its original form is not suited for multicultural assessment. Dana (2005) has provided information regarding usage of the TAT, taking cultural factors into consideration.

In addition, the TAT has been adapted for use with multicultural populations. Some adaptations have included changing the skin color of the individuals in photographs from White to Black as well as the creation of entirely separate tests, such as the Black Thematic Apperception Test (Bailey & Green, 1977) or Themes Concerning Blacks (Williams & Johnson, 1981). The Tell Me A Story Test is a culturally based narrative projective measure with a similar format to the TAT (i.e., pictorially based), originally developed for use with Spanish-speaking populations of children and adults, and includes culture-specific norms for Black, Puerto Rican, Hispanic and White children (Constantino & Malgady, 2000; Flanagan, Costantino, Cardalda, & Costantino, 2008). The TAT also has inspired other

versions of the test aimed at children, including the Children's Apperception Test among others (Groth-Marnat, 2009).

The TAT should be administered with caution to multicultural groups. Whenever possible, adaptations to the test for a specific group (e.g., Chinese) should be used, and clinicians should be aware that there are many different ways in which cultures express personality. In addition, clinicians should be aware of the effect of acculturation on TAT findings. The less acculturated a test taker is, the more caution a clinician needs to take when choosing to administer the TAT or to interpret its results. Clinicians must question whether TAT narratives are the result of personality structure or simply a reflection of cultural belief systems (Gray-Little, 2009; Groth-Marnat, 2009).

The Rorschach

There exists a history of controversy surrounding the use of the Rorschach, some of which is related to cross-cultural use. Although once hotly debated, the bulk of contemporary research now uses Exner's (2003) Comprehensive System for scoring. In terms of cross-cultural research, disagreement in research findings makes for difficult interpretation of cross-cultural validity across research articles (Allen & Dana, 2004).

The Rorschach does not include separate minority group norms, yet it is widely used with different ethnic groups. Some researchers have recommended that the test not be used for ethnic minority groups in the United States (Garb, Wood, Nezworski, Grove, & Stejskal, 2001), and others have reported that authors have misquoted their research to state their claim, which "obfuscates the intentions and accomplishments of current research efforts to develop an empirical basis for cross-cultural and multicultural Rorschach practice" (Allen & Dana, 2004, p. 191).

There are studies from earlier decades that revealed no difference between groups including Japanese Americans (e.g. Caudill, 1952, as cited in Ritzler, 2001) and African Americans in a meta-analytic study (Frank, 1992, as cited in Ritzler, 2001). Presley, Smith, Hilsenroth, and Exner (2001) found that a matched sample of 44 African Americans and 44 Caucasians differed only in their responses to Rorschach items on one variable, *S*, and in their significantly fewer Cooperative Movement responses. Some later writers have attributed the latter finding to other variables including reluctance of the African American sample to relate to Caucasian examiners (Ritzler, 2001).

In other research, no association between ethnicity and summary scores was found when looking at 432 Rorschach protocols (Meyer, 2002). This type of evidence bolsters the argument that there is no need for separate norming groups, as various ethnic groups do not appear to respond differently to the Rorschach (Butcher, 2009, p. 291). To further the conversation of ethnic group norms, continued research on specific groups should be pursued. Arguably, some studies (e.g., Meyer, 2002; Presley et al., 2001) have begun this process. Last, although group-specific research cannot be generalized to broader populations, there is some evidence that cultural minority or economically disadvantaged individuals score in a "less adaptive direction" on a Comprehensive System scoring of the Rorschach (Ritzler, 2001, p. 242).

In the past decade, Rorschach research has crossed international borders in non-U.S. countries (Weiner & Meyer, 2009). In 2007, the *Journal of Personality Assessment* devoted a special supplement issue on, "International Reference Samples for the Rorschach Comprehensive System," based on data from 17 different countries (Shaffer, Erdberg, & Meyer, 2007). Integrating findings from other internationally conducted research, the general conclusions of the studies included in the supplement indicated that adult populations are similar in their protocols. Authors recommended incorporating international composite reference values as a way of using the Rorschach in other countries with adults (Meyer, Erdberg, & Shaffer, 2007).

CULTURAL CONSIDERATIONS IN INTELLIGENCE TESTING

One of the most important components of any psychological evaluation is the assessment of intelligence. Intelligence "involves the ability to reason, plan, solve problems, think abstractly, comprehend

complex ideas, learn quickly, and learn from experience" ("Mainstream Science on Intelligence," 1994). However, culture determines to a large extent what is considered to be "intelligent" (Sternberg & Kaufman, 1998).

Intelligence testing has spurred a contentious cultural debate, given consistent findings of racial and ethnic group differences on the most frequently used measures. On the basis of a mean of 100 points (and $SD = 15$ points), a hierarchy of average scores by group are noted as follows: Whites, 100; Blacks, 85; Hispanics, midway between Blacks and Whites; and Asians and Jews somewhere above 100 ("Mainstream Science," 1994). Although the causes of such group differences (i.e., nature or nurture) continue to be debated in the literature, the measurement of intelligence has continued and grown in popularity with new theoretical and instrument development.

Closer examination of subtest scores on these intelligence tests indicates that groups also differ in terms of their overall profile of scores. For example, American Indian and Hispanic groups score relatively higher on visual reasoning in comparison with verbal reasoning tasks, and Asians tend to score higher in numerical and visual–spatial tasks (Suzuki, Vraniak, & Kugler, 1996). Recent editions of intelligence tests, including the Wechsler scales, continue to report group discrepancies, although differences have decreased over the years (e.g., Nisbett, 2009). Controversies have emerged regarding the role that intelligence tests play in the placement of students in special education (e.g., mental retardation, learning disabilities). Legal challenges have arisen, given disproportionately higher numbers of African American students being placed in classrooms for the mentally retarded in the 1960s and currently in programs that provide services for students with learning disabilities (e.g., Suzuki, Short, & Lee, 2011; see Volume 3, Chapter 3, this handbook, for more information on intelligence testing with children).

CULTURAL ASSESSMENT OF INTELLIGENCE

Efforts have been made to adjust scores on intelligence tests based on cultural factors. One of the first was the System of Multicultural Pluralistic Assessment (SOMPA; Mercer, 1979). The goal of the SOMPA was to modify scores on the Wechsler Intelligence Scale for Children—Revised (WISC–R) on the basis of age and sociocultural background. Other practices have included the Biocultural Model of Assessment (Armour-Thomas & Gopaul-McNicol, 1998) that attempted to integrate qualitative information about the examinee into the understanding of abilities utilizing preassessment information in the areas of health, language, previous experiences (educational and psychosocial), and family. The model incorporates biologically and culturally based instruments. Although the SOMPA and Biocultural models attempted to address the discrepancy between scores by race and ethnicity, the outcomes were discouraging. For example, the modified scores of the SOMPA predicted achievement less well than the actual WISC–R scores (Figueroa & Sassenrath, 1989). Partially because of these problems, implementation of these procedures did not become part of mainstream psychological assessment practices.

The Gf-Gc Cross-Battery Assessment Model (Flanagan, Ortiz, & Alfonso, 2007) is a promising method of intelligence assessment enabling clinicians to select among measures addressing broad and narrow ability areas. This method is based on an understanding of the cultural content and linguistic demands of different tests presented in the Culture–Language Test Classification (C-LTC; McGrew & Flanagan, 1998). The foundation of the C-LTC is the analysis of the cultural loading and linguistic demand for each classified measure. Classification is based on empirical test data, and expert consensus procedures are presented in the C-LTC.

Nonverbal measures of intelligence were developed in part to address the cultural loading present in language-based tests. Nonverbal tests include the Leiter International Performance Scale—Revised (Roid & Miller, 1997); Naglieri Nonverbal Ability Test (Naglieri, 1997); and the Universal Nonverbal Intelligence Test—2 (Bracken, Keith, & Walker, 1998). To solve the items presented on these measures, however, some degree of language or communication is involved, and the test items contain cultural information (Mpofu & Ortiz, 2009). Group differences continue to be found even on

these measures; thus, researchers refer to them as "culturally reduced" tests. It should be noted that intelligence is considered a multifactorial construct; therefore, focusing only on nonverbal areas of ability has limitations, given a focus on visual processing, memory, and speed of processing (Mpofu & Ortiz, 2009; see also Volume 3, Chapter 4, this handbook).

The Wechsler scales have maintained a monopoly in terms of popularity and frequency of use for decades. In their review of 59 intelligence/aptitude measures, Valencia and Suzuki (2001) noted that 66% of these studies used one or more of the Wechsler scales as a criterion measure. State-of-the-art intelligence tests such as the Wechsler scales use a number of procedures to address possible cultural bias, including expert reviews of content, racial and ethnic group oversampling, and specific statistical examination of scoring patterns by race and ethnic group. In addition, the Wechsler scales have been restandardized and renormed in several different countries. For example, the Wechsler Intelligence Scale for Children—III was standardized in the United Kingdom, France, Belgium, Germany, Sweden, Austria, Switzerland, Lithuania, Slovenia, Greece, Japan, South Korea, and Taiwan (Georgas, Weiss, van de Vijver, & Saklofske, 2003).

CONCLUSION

For decades, we have struggled as professionals to understand best practices in creating culturally responsive psychological assessment. Many scholars are quick to take a stance on the topic of multicultural assessment and split the practice of testing into discrete categories such as fair or unfair, groups into either marginalized or included, and results of testing into either useful or irrelevant. Despite posturing in discussions of educational and psychological assessment, what remains is the current and continued importance of clinical assessment in many spheres of our profession. Clinicians depend on psychological assessment to inform diagnosis and treatment planning. With this in mind, the authors reiterate the importance of knowledge, awareness, and skills to address the cultural complexities of the clinical assessment process.

The call for cultural competence in all areas of psychological practice requires that practitioners continue to examine the effect of culture on the performance outcomes of members of diverse communities. Although the focus of research has been on race and ethnic group differences, many clinical practitioners recognize this point as a limitation, given the importance of intersecting identities, which remains largely unexamined. Important aspects of identity may include region, language, religious beliefs, sexual orientation, social class, age, and ability status.

Although the psychometric properties of improved and technical procedures are readily used on most published measures, it is important to continue to be vigilant in and to continue to demand greater attention to diversity, including refugee and immigrant groups. In addition, training programs must include attention to cultural issues in assessment, the usage of translated versions of measures, and appropriate understanding of the role of interpreters.

Capturing the complexities of cultural context in assessment leads to the realization that definitive rules regarding appropriate assessment practices are perhaps impossible to attain. There are no culture-fair measures, and issues of equivalence will always be present. Clinicians must be ever watchful to ensure that we meet the assessment needs of our growing diverse clientele.

References

Allen, J., & Dana, R. H. (2004). Methodological issues in cross-cultural and multicultural Rorschach research. *Journal of Personality Assessment, 82*, 189–208. doi:10.1207/s15327752jpa8202_7

American Educational Research Association, American Psychological Association, & the National Council on Measurement and Evaluation. (2012). *Standards for educational and psychological testing*. Retrieved from http://teststandards.org

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

American Psychological Association. (n.d.). *Socioeconomic status*. Retrieved from http://www.apa.org/topics/socioeconomic-status/index.aspx

American Psychological Association. (2002). *Guidelines on multicultural education, training, research, practice, and organizational change for psychologists*.

Washington, DC: Author. Retrieved from http://www.apa.org/pi/oema/resources/policy/multicultural-guidelines.aspx

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Washington, DC: Author. Retrieved from http://www.apa.org/ethics/code/index.aspx

Arbisi, P. A., Ben-Porath, Y. S., & McNulty, J. (2002). A comparison of MMPI-2 validity in African American and Caucasian psychiatric inpatients. *Psychological Assessment, 14*, 3–15. doi:10.1037/1040-3590.14.1.3

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94. doi:10.1207/s15327752jpa8701_07

Armour-Thomas, E., & Gopaul-McNicol, S. (1998). *Assessing intelligence: Applying a bio-cultural model.* Thousand Oaks, CA: Sage.

Bailey, B. E., & Green, J. (1977). Black Thematic Apperception Test stimulus material. *Journal of Personality Assessment, 14*, 25–30.

Benson, E. (2003, February 3). Intelligent intelligence testing. *Monitor on Psychology, 34*(2), 48. Retrieved from http://www.apa.org/monitor/feb03/intelligent.aspx

Berg-Cross, L., & Takushi-Chinen, R. (1995). Multicultural training models and the Person-in-Culture Interview. In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (pp. 333–356). Thousand Oaks, CA: Sage.

Berry, J. W. (1989). Imposed etics-emics-derived etics: The operationalization of a compelling idea. *International Journal of Psychology, 24*, 721–735.

Berry, J. W. (2003). Conceptual approaches to acculturation. In K. Chun, P. Balls-Organista, & G. Marin (Eds.), *Acculturation: Advances in theory, measurement, and applied research* (pp. 17–37). Washington, DC: American Psychological Association.

Boring, E. G. (1923, June 6). Intelligence as the tests test it. *New Republic,* 35–37.

Bracken, B. A., Keith, L. K., & Walker, C. (1998). *Universal nonverbal intelligence test.* Itasca, IL: Riverside.

Butcher, J. N. (2000). Dynamics of personality test responses: The empiricist's manifesto revisited. *Journal of Clinical Psychology, 56*, 375–386. doi:10.1002/(SICI)1097-4679(200003)56:3<375::AID-JCLP13>3.0.CO;2-W

Butcher, J. N. (2009). *Oxford handbook of personality assessment.* New York, NY: Oxford University Press.

Butcher, J. N., & Ben-Porath, Y. S. (2004). Use of the MMPI–2 in medico-legal evaluations: An alternative interpretation for the Senior and Douglas critique. *Australian Psychologist*, *39*, 44–50. doi:10.1080/00050060410001660335

Butcher, J. N., Cheung, F. M., & Kim, J. (2003). Use of the MMPI–2 with Asian populations. *Psychological Assessment, 15*, 248–256. doi:10.1037/1040-3590.15.3.248

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi:10.1037/0735-7028.31.2.141

Castro, Y., Gordon, K. H., Brown, J. S., Anestis, J. C., & Joiner, T. E., Jr. (2008). Examination of racial differences on the MMPI–2 Clinical and Restructured Clinical Scales in an outpatient sample. *Assessment, 15*, 277–286. doi:10.1177/1073191107312735

Cheung, F. M., & Cheung, S. F. (2003). Measuring personality and values across cultures: Imported versus indigenous measures. In W. J. Lonner, D. L. Dinnel, S. A. Hayes, & D. N. Sattler (Eds.), *Online readings in psychology and culture* (Unit 6, Chapter 5). Bellingham: Center for Cross-Cultural Research, Western Washington University. Retrieved from http://www.wwu.edu/~culture

Cheung, F. M., Zhao, J. C., & Wu, C. Y. (1992). Chinese MMPI profiles among neurotic patients. *Psychological Assessment, 4*, 214–218. doi:10.1037/1040-3590.4.2.214

Choca, J. P., Shanley, L. A., Peterson, C. A., & Van Denberg, E. (1990). Racial bias and the MCMI. *Journal of Personality Assessment, 54*, 479–490.

Constantino, G., & Malgady, R. G. (2000). Multicultural and cross-cultural utility of the TEMAS (Tell-Me-a-Story) Test. In R. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 481–513). Mahwah, NJ: Erlbaum.

Dana, R. H. (1993). *Multicultural assessment: Perspectives for professional psychology.* Needham Heights, MA: Allyn & Bacon.

Dana, R. H. (2000). *Handbook of cross-cultural and multicultural personality assessment.* Mahwah, NJ: Erlbaum.

Dana, R. H. (2005). *Multicultural assessment: Principles, application, and examples.* Mahwah, NJ: Erlbaum.

Dillon, S. (2010, December 7). Top test scores from Shanghai stun educators. *New York Times.* Retrieved from http://www.nytimes.com/2010/12/07/education/07education.html

Ensminger, M. E., & Fothergill, K. E. (2003). A decade of measuring SES: What it tells us and where to go from here. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 13–27). Mahwah, NJ: Erlbaum.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*, 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Exner, J. E. (2003). *The Rorschach: A comprehensive system*. Hoboken, NJ: Wiley.

Figueroa, R. A., & Sassenrath, J. M. (1989). A longitudinal study of the predictive validity of the System of Multicultural Pluralistic Assessment (SOMPA). *Psychology in the Schools, 26*, 5–19. doi:10.1002/1520-6807(198901)26:1<5::AID-PITS2310260102>3.0.CO;2-D

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). San Francisco, CA: Wiley.

Flanagan, R., Costantino, G., Cardalda, E., & Costantino, E. (2008). TEMAS: A multicultural test and its place in an assessment battery. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed., pp. 323–345). San Francisco, CA: Wiley.

Flores, Y. L., & Obasi, E. M. (2003). Positive psychological assessment in an increasingly diverse world. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook* (pp. 41–54). Washington, DC: American Psychological Association. doi:10.1037/10612-003

Fuligni, A. J., & Yoshikawa, H. (2003). Socioeconomic resources, parenting, and child development among immigrant families. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 107–124). Mahwah, NJ: Erlbaum.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13*, 433–448. doi:10.1037/1040-3590.13.4.433

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (Eds.). (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. New York, NY: Academic Press.

Glass, M. H., Bieber, S. L., & Tkachuk, M. J. (1996). Personality styles and dynamics of Alaska Native and non-native incarcerated men. *Journal of Personality Assessment, 66*, 583–603. doi:10.1207/s15327752jpa6603_8

Goldberger, N. R., & Veroff, J. B. (Eds.). (1995). *The culture and psychology reader*. New York, NY: New York University Press.

Gray-Little, B. (2009). The assessment of psychopathology in racial and ethnic minorities. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 396–414). New York, NY: Oxford University Press.

Greene, R. L. (1987). Ethnicity and MMPI performance: A review. *Journal of Consulting and Clinical Psychology, 55*, 497–512. doi:10.1037/0022-006X.55.4.497

Greene, R. L., Robin, R. W., Albaugh, B., Caldwell, A., & Goldman, D. (2003). Use of the MMPI–2 in American Indians: II. Empirical correlates. *Psychological Assessment, 15*, 360–369. doi:10.1037/1040-3590.15.3.360

Grieger, I. (2008). A cultural assessment framework and interview protocol. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed., pp. 132–161). San Francisco, CA: Jossey-Bass.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Guerrero, M., & Heller, P. L. (2003). Sociocultural limits in informed consent in dementia research. *Alzheimer Disease and Associated Disorders*, *17*(Suppl. 1), 26–30. doi:10.1097/00002093-200304001-00005

Hahn, J. (2005). Faking bad and faking good by college students on the Korean MMPI–2. *Journal of Personality Assessment, 85*, 65–73. doi:10.1207/s15327752jpa8501_06

Hall, G. C. N., Bansal, A., & Lopez, I. R. (1999). Ethnicity and psychopathology: A meta-analytic review of 31 years of comparative MMPI/MMPI–2 research. *Psychological Assessment, 11*, 186–197. doi:10.1037/1040-3590.11.2.186

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Erlbaum.

Hill, J. S., Pace, T. M., & Robbins, R. R. (2010). Decolonizing personality assessment and honoring indigenous voices: A critical examination of the MMPI–2. *Cultural Diversity and Ethnic Minority Psychology*, *16*, 16–25. doi:10.1037/a0016110

Hill, J. S., Robbins, R. R., & Pace, T. M. (2012). Cultural validity of MMPI–2 empirical correlates: Is this the best we can do? *Journal of Multicultural Counseling and Development, 40*, 104–116. doi:10.1002/j.2161-1912.2012.00010.x

Hong, G. Y. (1998). Logistics and researchers as legitimate tools for "doing" intercultural research: A rejoinder to Günther. *Culture and Psychology, 4*, 81–90. doi:10.1177/1354067X9800400106

Hoop, J. G., DiPasquale, T., Hernandez, J. M., & Roberts, L. W. (2008). Ethics and culture in mental health

care. *Ethics and Behavior, 18*, 353–372. doi:10.1080/10508420701713048

Intervention Central. (2011). *CBM maze passage generator*. Retrieved from http://www.interventioncentral.com

Ketterer, H. L., Han, K., Hur, J., & Moon, K. (2010). Development and validation of culture-specific Variable Response Inconsistency and True Response Inconsistency Scales for use with the Korean MMPI–2. *Psychological Assessment, 22*, 504–519. doi:10.1037/a0019511

Kim, B. S. K., & Abreu, J. M. (2001). Acculturation measurement theory, current instruments, and future directions. In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (2nd ed., pp. 394–424). Thousand Oaks, CA: Sage.

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Mivile, M. L., Kerns, R., & Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology, 60*, 725–739. doi:10.1002/jclp.20010

Kroeber, A. L., & Kluckhohn, C. (1963). *Culture: A critical review of concepts and definitions.* Cambridge, MA: Harvard University Press.

Kwan, K. L. K., & Maestas, M. L. (2008). MMPI-2 and MCMI-III performances of non-White people in the United States: What we (don't) know and where we go from here. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (3rd ed., pp. 425–446). San Francisco, CA: Jossey-Bass.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.

Lopez, E. C. (2010). Interpreters. In C. C. Clauss-Ehlers (Ed.), *Encyclopedia of cross-cultural school psychology* (pp. 547–553). New York, NY: Springer.

López, S. R., & Guarnaccia, J. J. (2000). Cultural psychopathology: Uncovering the social world of mental illness. *Annual Review of Psychology, 51*, 571–598. doi:10.1146/annurev.psych.51.1.571

Mainstream science on intelligence. (1994, December 13). *Wall Street Journal,* A18.

McGrew, K. S., & Flanagan, D. P. (1998). *The Intelligence Test Desk Reference (ITDR): Gf-Gc cross-battery assessment*. Boston, MA: Allyn & Bacon.

McShane, D. (1980). A review of scores of American Indian children on the Wechsler Intelligence Scale. *White Cloud Journal, 2*, 18–22.

Mercer, J. (1979). In defense of racially and culturally non-discriminatory assessment. *School Psychology Review, 8*, 89–115.

Meyer, G. J. (2002). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment, 78*, 104–129. doi:10.1207/S15327752JPA7801_07

Meyer, G. J., Erdberg, P., & Shaffer, T. W. (2007). Toward international normative reference data for the Comprehensive System. *Journal of Personality Assessment, 89*(Suppl. 1), 201–216.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2006). *Millon Clinical Multiaxial Inventory—III manual* (3rd ed.). Minneapolis, MN: Pearson.

Monopoli, J., & Alworth, L. (2000). The use of the thematic apperception test in the study of Native Americans' psychological characteristics: A review and archival study of Navaho men. *Genetic, Social, and General Psychology Monographs, 126*, 43–78.

Mpofu, E., & Ortiz, S. O. (2009). Equitable assessment practices in diverse contexts. In E. L. Grigorenko (Ed.), *Multicultural psychoeducational assessment* (pp. 41–76). New York, NY: Springer.

Munley, P. H., Vacha-Haase, T., Busby, R. M., & Paul, B. D. (1998). The MCMI–II and race. *Journal of Personality Assessment, 70*, 183–189. doi:10.1207/s15327752jpa7001_12

Murray, H. A. (1943). *The Thematic Apperception Test*. Cambridge, MA: Harvard University Press.

Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. Upper Saddle River, NJ: Pearson.

National Council on Interpreting in Health Care. (2004). *National code of ethics for interpreters in health care*. Retrieved from http://www.ncihc.org/mc/page.do?sitepages

National Council on Interpreting in Health Care. (2005). *National standards of practice for interpreters in health care*. Retrieved from http://www.ncihc.org/mc/page.do?sitepages

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. New York, NY: Norton.

Oakland, T. (2009). How universal are test development and use? In E. L. Grigorenko (Ed.), *Multicultural psychoeducational assessment* (pp. 1–40). New York, NY: Springer.

Ochoa, S. H. (2003). Assessment of culturally and linguistically diverse children. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 563–583). New York, NY: Guilford Press.

Organisation for Economic Co-operation and Development. (2009). *Program for international student assessment: Figure 1: Comparing countries' and economies' performance*. Retrieved from http://www.oecd.org/pisa/46643496.pdf

Ortiz, S. O., & Dynda, A. M. (2005). Use of intelligence tests with culturally and linguistically diverse populations. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (2nd ed., pp. 545–556). New York, NY: Guilford Press.

Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 234–250). New York, NY: Guilford Press.

Pace, T. M., Robbins, R. R., Choney, S. K., Hill, J. S., Lacey, K., & Blair, G. (2006). A cultural-contextual perspective on the validity of the MMPI–2 with American Indians. *Cultural Diversity and Ethnic Minority Psychology, 12*, 320–333. doi:10.1037/1099-9809.12.2.320

Pedersen, P. (Ed.). (1999). Culture-centered interventions as a fourth dimension in psychology. In *Multiculturalism as a fourth force* (pp. 3–18). New York, NY: Sage.

Piekkola, B. (2011). Traits across cultures: A neo-Allportian perspective. *Journal of Theoretical and Philosophical Psychology, 31*, 2–24. doi:10.1037/a0022478

Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical training programs. *Journal of Personality Assessment, 61*, 394–405. doi:10.1207/s15327752jpa6102_17

Presley, G., Smith, C., Hilsenroth, M., & Exner, J. (2001). Clinical utility of the Rorschach with African Americans. *Journal of Personality Assessment, 77*, 491–507. doi:10.1207/S15327752JPA7703_09

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 332–374). Hoboken, NJ: Wiley.

Ritzler, B. A. (2001). Multicultural usage of the Rorschach. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (2nd ed., pp. 237–252). San Francisco, CA: Jossey Bass.

Rivera, L. M. (2008). Acculturation and multicultural assessment: Issues, trends, and practice. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment* (3rd ed., pp. 73–91). San Francisco, CA: Jossey-Bass.

Robin, R. W., Greene, R. L., Albaugh, B., Caldwell, A., & Goldman, D. (2003). Use of the MMPI–2 in American Indians: I. Comparability of the MMPI–2 between two tribes and with the MMPI–2 normative group. *Psychological Assessment, 15*, 351–359. doi:10.1037/1040-3590.15.3.351

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale—Revised.* Los Angeles, CA: Western Psychological Services.

Sattler, J. M., & Hoge, R. D. (2006). *Assessment of children: Behavioral, social and clinical foundations* (5th ed.). San Diego, CA: Jerome M. Sattler.

Serpell, R. (2000). Intelligence and culture. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 549–577). New York, NY: Cambridge University Press.

Shaffer, T. W., Erdberg, P., & Meyer, G. J. (2007). Introduction to the JPA special supplement on international reference sampled for the Rorschach Comprehensive System. *Journal of Personality Assessment, 89*, S2–S6. doi:10.1080/00223890701629268

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49*, 479–502. doi:10.1146/annurev.psych.49.1.479

Sue, D. W., & Sue, D. (2008). *Counseling the culturally diverse: Theory and practice* (5th ed.). Hoboken, NJ: Wiley.

Sue, S., Keefe, K., Enomoto, K., Durvasula, R., & Chao, R. (1996). Asian American and White college students' performance on the MMPI–2. In J. N. Butcher (Ed.), *International adaptations of the MMPI Research and clinical applications* (pp. 206–220). Minneapolis: University of Minnesota Press.

Sukigara, M. (1996). Equivalence between computer and booklet administrations of the new Japanese version of the MMPI. *Educational and Psychological Measurement, 56*, 570–584. doi:10.1177/0013164496056004002

Suzuki, L. A., Short, E. S., & Lee, C. S. (2011). Racial and ethnic group differences in intelligence in the United States. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 273–292). New York, NY: Cambridge University Press.

Suzuki, L. A., Vraniak, D. A., & Kugler, J. F. (1996). Intellectual assessment across cultures. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 141–178). San Francisco, CA: Jossey Bass.

Timbrook, R. E., & Graham, J. R. (1994). Ethnic differences on the MMPI–2? *Psychological Assessment, 6*, 212–217. doi:10.1037/1040-3590.6.3.212

Tsai, D. C., & Pike, P. L. (2000). Effects of acculturation on the MMPI–2 scores of Asian American students. *Journal of Personality Assessment, 74*, 216–230. doi:10.1207/S15327752JPA7402_4

Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues.* Thousand Oaks, CA: Sage.

Valencia, R. R., Suzuki, L. A., & Salinas, M. F. (2000). Test bias. In R. R. Valencia & L. A. Suzuki (Eds.), *Intelligence testing and minority students: Foundations,*

performance factors, and assessment issues (pp. 111–150). Thousand Oaks, CA: Sage.

Velasquez, R. J., Gonzales, M., Butcher, J. N., Castillo-Canez, I., Apodaca, J. X., & Chavira, D. (1997). Use of the MMPI–2 with Chicanos: Strategies for counselors. *Journal of Multicultural Counseling and Development, 25*, 107–120. doi:10.1002/j.2161-1912.1997.tb00321.x

Waggoner, W. C., & Mayo, D. M. (1995). Who understands? A survey of 25 words or phrases commonly used in proposed clinical research consent forms. *IRB: Ethics and Human Research*, *17*, 6–9. doi:10.2307/3563639

Weiner, I. B., & Meyer, G. J. (2009). Personality assessment with the Rorschach Inkbot Method. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 277–298). New York, NY: Oxford University Press.

Wetzler, S. (1990). The Millon Clinical Multiaxial Inventory (MCMI): A review. *Journal of Personality Assessment, 55*, 445–464.

Whatley, P. R., Allen, J., & Dana, R. H. (2003). Racial identity and the MMPI in African American male college students. *Cultural Diversity and Ethnic Minority Psychology, 9*, 345–353. doi:10.1037/1099-9809.9.4.345

Williams, R. L., & Johnson, R. C. (1981). Progress in developing Afrocentric measuring instruments. *Journal of Non-White Concerns, 10*, 3–18. doi:10.1002/j.2164-4950.1981.tb00045.x

CHAPTER 13

PSYCHOLOGICAL ASSESSMENT IN TREATMENT

Michael J. Lambert and David A. Vermeersch

Psychological testing has its roots in three earlier developments: civil service examinations, the assessment of academic achievement in universities and schools, and studies by European and American scientists on the measurement of individual differences in behavior (DuBois, 1970). Regardless of the context in which a test was used, the primary purpose for using it was to identify, illuminate, and explore differences that existed between examinees. That is, the need to differentiate among individuals was the theoretical basis for virtually all of the early efforts in psychological testing (Cronbach, 1984). The early emphasis placed on the study of individual differences has strongly influenced modern psychological testing and measurement, which originated in investigations of person-to-person variability in functions such as sensory discrimination, reaction time, perceptual abilities, motor skills, and problem solving (DuBois, 1970). Currently, the identification and exploration of individual differences continues to be the primary purpose of the vast majority of available psychological tests. However, the use of measures designed to assess individual change over time has increased dramatically as a function of the widespread interest in measuring the effects of psychotherapy.

Kirshner and Guyatt (1985) long ago noted that of the thousands of psychological tests that have been published to date, most have been specifically designed to serve one or more of the following purposes: discrimination, prediction, and evaluation. A discriminative measure is one that is used for the purpose of distinguishing between or among individuals or groups on the basis of an underlying dimension when no external criterion or gold standard is available. Intelligence tests such as the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV; Wechsler, 2008) and personality inventories such as the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher et al., 2001) are examples of discriminative measures. Discriminative measures are often used as diagnostic instruments because they are specifically designed to discriminate among different individuals (based on their scores on the respective measure) at a single point in time.

A predictive measure is one that is used for the purpose of classifying individuals into categories when a gold standard is available—either concurrently or prospectively—that can determine whether individuals have been classified correctly. This type of measure is generally used as a screening instrument to identify which specific individuals have or will develop a target condition. When a measure is used to assist in assessing whether a patient is appropriate for a specific type of treatment (e.g., using a measure to assess a patient's ego strength for the purpose of predicting whether he or she will be able to meaningfully participate in brief dynamic therapy), that measure is being used for predictive purposes. The Child Abuse Potential Inventory (Milner, Gold, Ayoub, & Jacewitz, 1984) is an example of a predictive measure in that it is designed to detect individuals who are at an increased risk of committing abusive acts.

An evaluative measure is one that is used for the purpose of measuring change over time (e.g., pre- and

DOI: 10.1037/14048-013
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

posttreatment, weekly change over the course of treatment) in an individual or group on the dimension(s) of interest. Tests designed to assess treatment benefits or outcomes are examples of evaluative measures. Outcome measures, then, are quite different from discriminative and predictive measures because they are designed to measure intraindividual change over time through repeated administrations rather than discriminate between or among different individuals at a single time point (Kirshner & Guyatt, 1985).

Psychological measures are often used, either appropriately or inappropriately, for some combination of the aforementioned purposes. For example, a measure may be used to assist in determining a patient's appropriateness for a specific type of treatment (i.e., predictive purpose) and then subsequently used to track that patient's progress throughout the course of treatment and status at termination (i.e., evaluative purpose). Although this process is justified in cases in which a measure has demonstrated utility in serving multiple purposes, such a practice often represents an application of a measure for a purpose other than that for which the measure was designed and may lead to the inaccurate assessment of a patient. For example, Froyd, Lambert, and Froyd (1996) reported the MMPI to be among the 10 most frequently used self-report measures of outcome (i.e., evaluative purpose), despite the fact that the MMPI was originally designed for diagnostic purposes (i.e., discriminative purpose). For further discussion of the MMPI and other self-report measures of personality, please refer to the Chapter 11 in this volume. The MMPI is not an appropriate instrument for measuring outcome because it contains many items that are not sensitive to changes in patients receiving treatment, is excessively long, and is relatively expensive. As this example suggests, it is extremely important to take care when selecting instruments to measure patient response to treatment.

Although psychological assessment in treatment can involve the use of discriminative and predictive indexes, the primary focus of psychological assessment in treatment is the measurement of intraindividual change over time, or psychotherapy outcome, of patients receiving treatment. Therefore, the present chapter focuses on theoretical, conceptual, methodological, and applied issues relevant to psychological assessment in treatment and the establishment of a comprehensive system for measuring and improving psychotherapy outcome. Readers also may want to see Chapter 18, this volume, on outcomes assessment in health settings for related information.

TEST SELECTION IN PSYCHOTHERAPY OUTCOME

In measuring psychotherapy outcome, the ability to assess patient response to treatment accurately throughout the course of therapy, at termination, and at follow-up is directly related to the quality and appropriateness of the measure(s) being used for this purpose (Ogles, Lambert, & Fields, 2002). Therefore, it is imperative that appropriate measures of outcome are used, or else treatment gains may go undetected, a mistake that clinicians and researchers can ill afford to make in an age of increased accountability. Researchers and clinicians would therefore benefit from being aware of the qualities associated with sound outcome measures.

The development of selection criteria (i.e., characteristics of instruments that will lead to the most accurate measurement of patient change) for outcome measures to be implemented in practice has received increased attention in recent years (Trabin, 1995). Because professional practices are coming to rely heavily on the demonstration of measured effects of treatments, it is imperative that outcome measures possess characteristics that will lead to the most accurate reflection of patient improvement. Some authors (Horowitz, Milbrath, & Stinson, 1997; Pilkonis, 1997; Shea, 1997) have proposed selection criteria for instruments aimed at measuring changes in symptomatology associated with a specific disorder (e.g., major depressive disorder) or a major diagnostic category (e.g., personality disorders). Others have focused on the development of universally accepted selection criteria that can be applied to the evaluation of any outcome measure (Lambert, Horowitz, & Strupp, 1997; Newman & Ciarlo, 1994). Although there are some differences

between the selection criteria proposed by various authors, there appears to be considerable overlap as well.

Synthesizing and building upon the available literature, Lambert et al. (1997) suggested that the following 13 criteria consistently emerge as appropriate means of selecting methods and measures of outcome: (a) relevance to target group; (b) simple, teachable methods; (c) objective referents; (d) multiple respondents; (e) psychometric strengths and the availability of norms; (f) low measure costs relative to its use; (g) understanding by nonprofessional audiences, easy feedback, uncomplicated interpretation; (h) utility in clinical services; (i) compatibility with a variety of clinical theories and practices; (j) the possibility of multiple administrations; (k) comprehensiveness; (l) relationship to a diagnostic classification system (e.g., the *Diagnostic and Statistical Manual of Mental Disorders* [*DSM*]); and (m) sensitivity to change (i.e., the ability of an outcome measure to detect change following an intervention). Given that the central focus of psychotherapy outcome assessment is the detection of intraindividual change over time, the sheer importance of change sensitivity as a criterion in outcome test selection warrants special attention.

SENSITIVITY TO CHANGE

When patient changes are not detected on an outcome measure, it is likely that either the treatment did not work or the instrument was inadequate in detecting changes that occurred (Guyatt, 1988). In psychotherapy outcome assessment, the sensitivity to change of an outcome measure refers to the degree to which an instrument accurately reflects patient changes that occur following participation in therapy (Hill & Lambert, 2004). Therefore, the sensitivity to change of a measure is directly related to the construct validity of the instrument, because the primary purpose of outcome measures is to document patient changes after a course of therapy. Given the central importance of change sensitivity in outcome assessment, it is necessary to gather information about the sensitivity to change of an outcome measure before it can be confidently used to assess the effects of treatment on patients. However, rarely, if ever (at least at the time of first publication), do outcome test manuals contain information (e.g., repeated measures data on various patient and nonpatient samples) that supports the sensitivity to change of the items, subscales, and total score of a measure. In a review of 348 outcome studies utilizing 1,430 distinct measures, Froyd et al. (1996) found that virtually none of the top 10 most frequently used measures provide evidence in their test manuals supporting the sensitivity to change of the items, subscales, or total score of the test. Failure to gather information regarding the change sensitivity of measures may lead to inaccurate conclusions regarding the effectiveness of interventions.

Several researchers have developed methodologies that can be used in the evaluation of change sensitivity (Guyatt, 1988; Meier, 1997; Tryon, 1991; Vermeersch, Lambert, & Burlingame, 2000; Vermeersch et al., 2004). In an attempt to synthesize and build upon the literature on change sensitivity, Vermeersch et al. (2000, 2004) proposed two criteria for establishing the change sensitivity of an outcome measure: (a) patient change on an item, subscale, or total score of an outcome measure should occur in the theoretically proposed direction (i.e., most often, change reflective of patient improvement over the course of treatment); and (b) the change observed on an item, subscale, or total score of an outcome measure indicates significantly more improvement in treated than in untreated individuals. These researchers also applied this methodology to the items, subscales, and total score of a widely used outcome measure for the purpose of assessing the appropriateness of using the measure to assess outcome in specific patient populations. However, it is important to note that this methodology is not measure specific and can be applied to any measure that is used or may potentially be used to assess patient changes in psychotherapy.

Figures 13.1 through 13.4 are derived from the work of Vermeersch et al. (2000, 2004) on change sensitivity. The figures illustrate broad categorizations of patterns of change that can be observed on the items, subscales, and total scores of various outcome measures. These broad categorizations are offered as examples of various patterns of patient

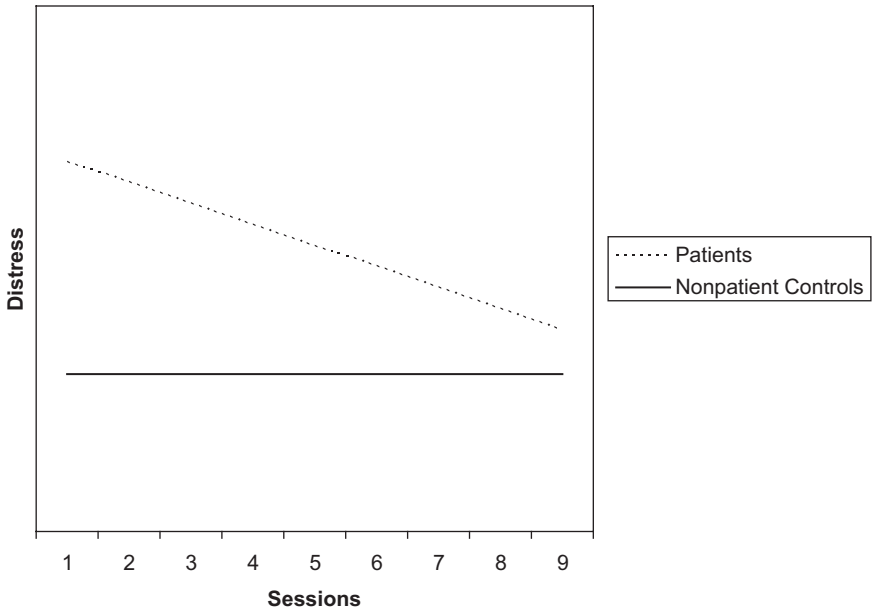


FIGURE 13.1. Example A, illustrating patterns of change indicative of change sensitivity in an outcome measure. The patient and nonpatient patterns of change suggest a measure that is sensitive to change (in that patients demonstrate change over time and these changes exceed those observed in nonpatient controls) and effective in detecting interindividual differences at pretreatment.

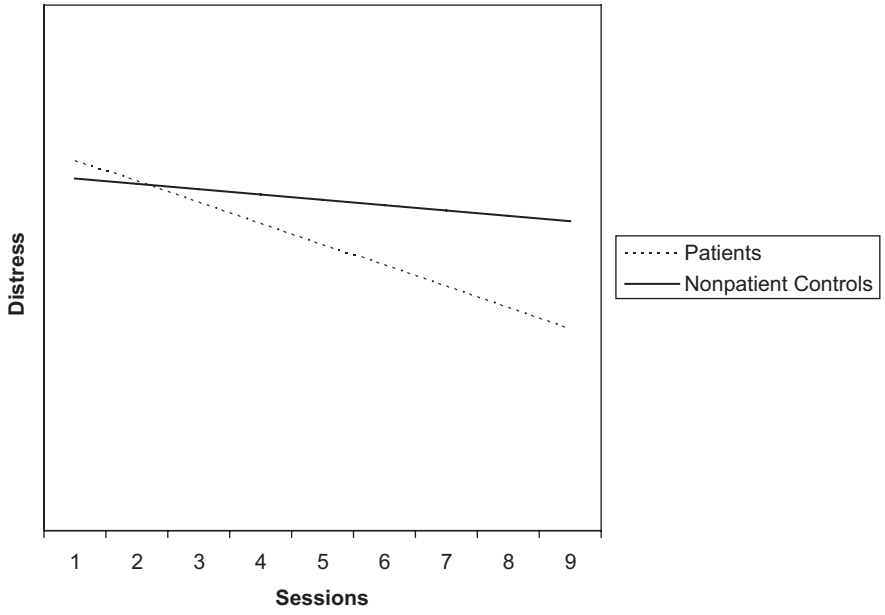


FIGURE 13.2. Example B, illustrating patterns of change indicative of change sensitivity in an outcome measure. The patient and nonpatient patterns of change suggest a measure that is sensitive to change (in that patients demonstrate change over time and these changes exceed those observed in nonpatient controls), despite the fact that the measure is not effective in detecting interindividual differences at pretreatment.

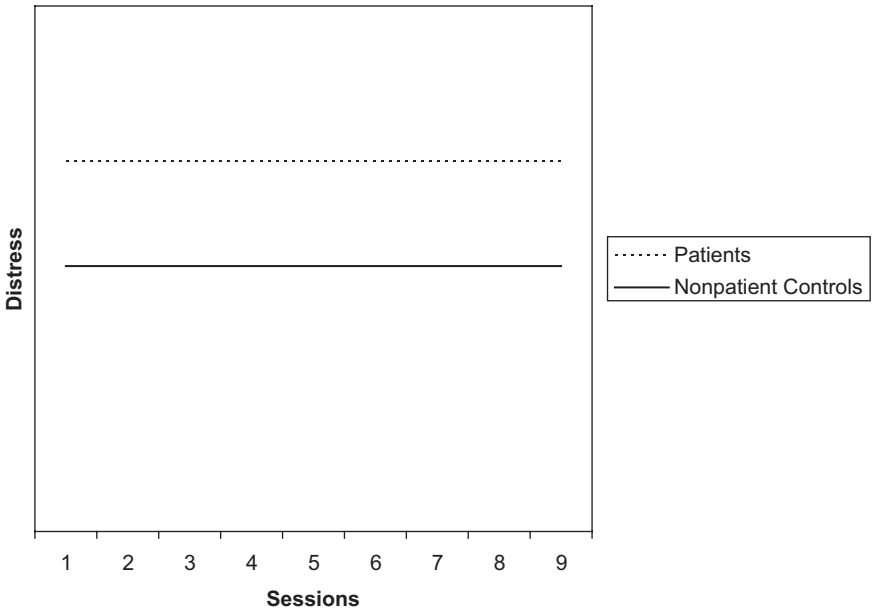


FIGURE 13.3. Example A, illustrating patterns of change indicative of lack of change sensitivity in an outcome measure. The patient and nonpatient patterns of change suggest a measure that is not sensitive to change in that patients do not demonstrate improvements and the change observed in patients does not exceed the change observed in nonpatient controls. Note that this measure does detect interindividual differences between patients and nonpatients at pretreatment.

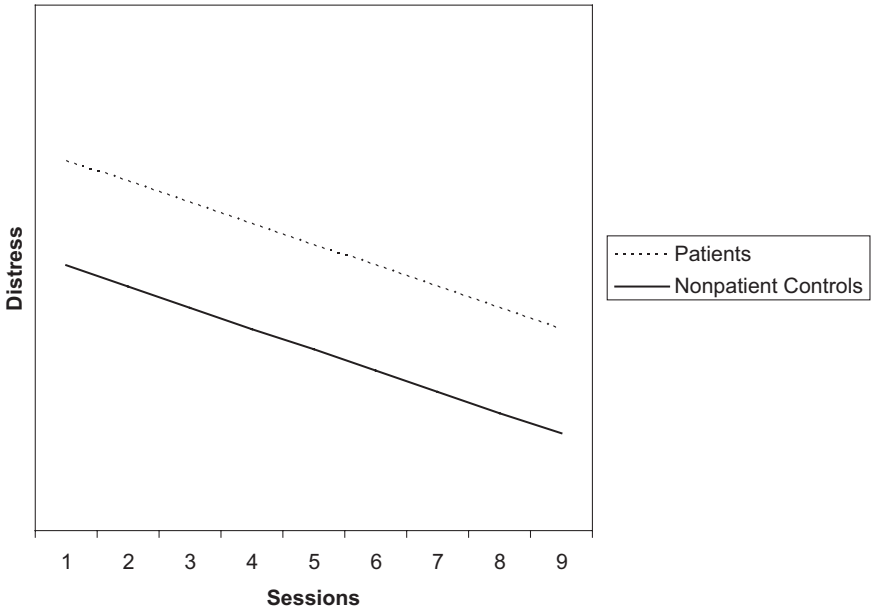


FIGURE 13.4. Example B, illustrating patterns of change indicative of lack of change sensitivity in an outcome measure. The patient and nonpatient patterns of change suggest a measure that detects patient changes. However, these changes do not exceed changes observed in nonpatient controls, suggesting that the measure lacks in change sensitivity. Note that this measure does detect interindividual differences between patients and nonpatients at pretreatment.

change indicative of change sensitivity and change insensitivity.

The patterns of change noted in Figure 13.1 demonstrate change sensitivity according to the aforementioned criteria in that patient change occurs in the theoretically proposed direction (i.e., improvement) and the change observed indicates significantly more improvement in treated than in untreated individuals. In addition to demonstrating intraindividual change over time, it is also important to note that this instrument appears to be able to detect interindividual differences at pretreatment (as evidenced by significantly higher scores before Session 1 for patients than nonpatients), suggesting that this measure could potentially be used for a discriminative purpose as well as in measuring outcome.

Although the patterns of change noted in Figure 13.2 suggest that the measure does not appear to be effective in detecting interindividual differences at pretreatment (as evidenced by relatively small differences between patients and nonpatient controls at pretreatment), the measure does demonstrate change sensitivity in that patient change occurs in the theoretically proposed direction and the change observed indicates significantly more improvement in treated than in untreated individuals. Although it is somewhat unlikely that a measure would demonstrate sensitivity to patient changes over time yet fail to detect interindividual differences at pretreatment, the pattern of change in Figure 13.2 is nonetheless suggestive of an instrument that could effectively be used to measure psychotherapy outcome.

In contrast to Figures 13.1 and 13.2, Figures 13.3 and 13.4 depict measures that are not sensitive to change. In Figure 13.3, the measure is able to detect differences between patients and nonpatients at pretreatment, but the measure does not demonstrate change over time, nor does it demonstrate significantly more change in patients than in nonpatient controls. This profile suggests that this instrument, although effective in detecting interindividual differences at a single point in time, is not effective in assessing the effects of therapy. Similarly, Figure 13.4 suggests that the subject measure is effective in detecting interindividual differences at pretreatment. Furthermore, the instrument meets one criterion of change sensitivity in that patients change in the theoretically proposed direction. However, this measure fails to demonstrate change sensitivity in that patient changes do not exceed those made by nonpatient controls over relatively brief periods of time.

It is important to note that in Figures 13.3 and 13.4, the measure was able to detect interindividual differences at a single point in time (i.e., pretreatment). However, placing primary emphasis on the detection of interindividual differences at a single time point in outcome test development (which is the current practice, given that traditional test development procedures are applied to the development of outcome measures) and outcome test selection (which is often the case when researchers and practitioners attempt to use measures that were designed to detect interindividual differences to measure intraindividual change) is problematic. Although the test developer or user will be able to use the measure to discriminate between different individuals at a single time point, it will be unknown whether the measure will achieve its primary purpose of detecting patient changes following intervention (as depicted in Figures 13.1 and 13.2) or will lack in change sensitivity (as depicted in Figures 13.3 and 13.4).

Although the aforementioned methodology offers a means of assessing change sensitivity, it does not ascertain the factors underlying lack of change sensitivity. First, measures may include items that are not relevant to the group under investigation, resulting in lack of measured change (Fitzpatrick et al., 1992). Second, measures that contain items that are categorically arranged (e.g., yes–no or true–false) or that offer only a limited range of response options (e.g., a Likert-type scale ranging from 0 to 2) may be scaled in units that are too gross to detect change (Lipsey, 1990). Third, measures may contain instructions that are not conducive to the detection of change. For example, an instrument that asks respondents to answer items according to how they have felt over an extended period of time (e.g., the past 3 months) is not likely to be useful in detecting changes resulting from interventions that have been delivered weekly over a brief period of time (Berrett, 1998). Fourth, measures may contain items that assess areas that are relatively stable or not a feasible

target of assessment (e.g., assessment of stable personality characteristics). Fifth, measures may contain items that are susceptible to floor or ceiling effects and, therefore, limit the ability of the item to detect meaningful growth or deterioration (Lipsey, 1990).

The importance of selecting an appropriate outcome measure(s) in light of the target patient population cannot be overemphasized, as the success of the entire outcome measurement system will be largely contingent on the measure used. Attention to the aforementioned criteria for outcome test selection, particularly the change sensitivity of a measure, will increase the likelihood of selecting a measure that provides the most accurate reflection of patient outcome. Accurate assessment of patient outcome will, in turn, allow clinicians to demonstrate treatment effectiveness more convincingly and alter treatment strategies if patients are nonresponsive to or deteriorating during treatment.

THE NEW STANDARD OF MANAGING OUTCOMES

A major emerging trend in psychotherapy outcome research is the shift from merely measuring and monitoring outcome to managing outcome (Lambert, 2010). For many decades, and even to the present day, psychotherapy outcome research (with the notable exception of the behavior therapies) has relied heavily on research designs that measure patient outcome at pre- and posttreatment. Although such designs have proven beneficial in establishing the general efficacy and effectiveness of the treatments under investigation, they are limited in that outcome data from these studies (because they are collected following termination from treatment) cannot be used to positively influence the treatment process of the individual patients under investigation. Pre- and posttreatment assessments, then, constitute essentially a "postmortem" analysis of outcome, because patients have already terminated treatment and nothing can be done to improve their outcomes, even if they experienced no change or deteriorated while in treatment.

A more recent trend in outcome research is the increased emphasis on regularly monitoring or tracking outcome throughout the treatment process. Regular monitoring of patient progress has, in addition to answering questions related to the general efficacy and effectiveness of treatments, allowed researchers to explore more sophisticated questions related to psychotherapy outcome. For example, through regularly monitoring patient change throughout treatment, researchers have been able to better understand patterns of change in psychotherapy, as evidenced in the growing and evolving body of literature related to the dose–response relationship in psychotherapy. However, researchers have not typically used data from studies involving the regular monitoring of patient progress in real time to influence treatment process and outcome positively, a pattern similar to that which has historically occurred in studies involving the pre- and posttreatment assessment of outcome. Although this issue does not pose a significant concern for the large number of patients who respond well to treatment and attain positive outcomes, it is particularly problematic for the large minority of patients who proceed completely through a course of treatment and experience no change (approximately 30%–40%) or actually deteriorate (5%–15%) during the course of therapy (Hansen, Lambert, & Forman, 2002; Lambert & Ogles, 2004).

Outcome management extends the practice of measuring and monitoring patient progress throughout the course of treatment by using these data (often in real time) to positively influence the treatment process and outcome of these same patients. The major advantage of psychotherapy outcome management is that outcome data can be regularly gathered and used by administrators and clinicians for the purpose of making needed alterations in intervention strategy if patients engaged in treatment are either unresponsive or deteriorating, which is a primary concern of virtually all stakeholders in the treatment process.

Several psychotherapy outcome management systems have been developed and implemented in clinical service delivery settings worldwide. Although a discussion of the specific procedures used in each of these quality management systems vary, a common feature across all of them is the monitoring of patient outcome throughout the

course of treatment and the use of these data to improve outcomes (Barkham et al., 2001; Howard, Kopta, Krause, & Orlinsky, 1986; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Kordy, Hannöver, & Richard, 2001; Kraus & Horan, 1997; Miller, Duncan, Sorrell, & Brown, 2005). However, conclusions about the relative value of each of these systems for enhancing patient outcome are still in question because very little research on the preceding systems has evaluated the effects of feedback on patient outcome using randomized controlled trials.

In the remainder of this chapter, one specific psychotherapy quality management system that has been developed, implemented, and empirically evaluated through multiple, randomized controlled trials (Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert, Whipple, Smart et al., 2001; Whipple et al., 2003) is described. Specifically, the major components of this system are provided as well as a description of how the provision of regular feedback to clinicians on their patients' progress has been used to improve outcomes, particularly for those patients who are not having a favorable response to therapy. Although the specific aspects of this psychotherapy quality management system are described, the concepts discussed are relevant to any researcher or clinician attempting to track patient change in treatment and ultimately use this information to improve outcomes.

THE OUTCOME QUESTIONNAIRE (OQ) PSYCHOTHERAPY QUALITY MANAGEMENT SYSTEM

The OQ

Given the demand for regular and efficient outcome assessment in psychotherapy outcome management, and in accordance with the aforementioned criteria for outcome test selection, a suitable measure was selected for implementation in the aforementioned psychotherapy quality management system. The OQ (Lambert, Morton, et al., 2004) is a 45-item, self-report measure designed for repeated administration throughout the course of treatment and at termination of treatment. Although the OQ can be administered and scored by hand in a short period of time (i.e., 5–7 minutes), it is ideally administered and scored electronically so that outcome results are immediately generated and made available to the treating clinician. In accordance with several reviews of the literature (e.g., Lambert, 1983), the OQ was conceptualized and designed to assess three domains of patient functioning: (a) symptoms of psychological disturbance (particularly anxiety and depression), (b) interpersonal problems, and (c) social role functioning. Consistent with this conceptualization of outcome, the OQ provides a Total Score, based on all 45 items as well as Symptom Distress, Interpersonal Relations, and Social Role subscale scores. Each of these subscales contains some items related to the quality of life of the individual. Higher scores on the OQ are indicative of greater levels of psychological disturbance.

Research has indicated that the OQ is a psychometrically sound instrument, with excellent internal consistency (coefficient alpha of .93 for the Total Score), adequate 3-week test–retest reliability (r = .84), and strong concurrent validity estimates ranging from .55 to .88 (all significant at $p < .01$) when the Total Score and subscale scores were correlated with scores from other widely used measures (Lambert, Morton, et al., 2004). Furthermore, the OQ has been shown to be sensitive to changes in multiple patient populations over short periods of time while remaining relatively stable in untreated individuals (Vermeersch et al., 2000, 2004). In short, the OQ is a brief measure of psychological disturbance that yields scores that are reliable and sensitive to changes patients make during psychotherapy; evidence of test score validity is compelling. It is well suited for tracking patient status during and after treatment.

Defining a Positive and Negative Outcome

A key element in assessing the effects of treatment is defining and operationalizing the concepts of positive and negative outcome. Jacobson and Truax (1991) offered a methodology by which patient changes on an outcome measure can be classified as recovered, reliably improved, showing no change, or deteriorated. There are two pieces of information necessary to make patient outcome classifications: a Reliable Change Index (RCI) and a normal functioning cutoff score. Clinical and normative data were

analyzed by Lambert, Morton, and colleagues (2004) to establish an RCI and a cutoff score for the OQ. The RCI obtained on the OQ was 14 points, indicating that patient changes of 14 or more points on the OQ can be considered reliable (i.e., not due to measurement error). The cutoff score for normal functioning on the OQ was calculated to be 63, indicating that scores of 64 or higher are more likely to come from a dysfunctional population than a functional population, and scores of 63 or lower are more likely to come from a functional population than a dysfunctional population. Using this information, patients can be placed in the following categories based on the change observed in their OQ scores: *Recovered* (i.e., clinically significant change) referred to patients whose scores decrease by 14 or more points and pass below the cutoff score of 64. *Improved* (i.e., reliably changed) referred to patients whose scores decrease by 14 or more points but do not pass below the cutoff score of 64. *No change* referred to patients whose scores change by less than 14 points in either direction. *Deteriorated* referred to patients whose scores increase by 14 or more points.

Support for the validity of the OQ's reliable change and cutoff score has been reported by Lunnen and Ogles (1998) and Beckstead et al. (2003). Having a method to classify each patient's treatment response is an essential component of any psychotherapy quality management system, given that the primary purpose of such systems is to understand and improve on the gains each individual makes during the course of treatment. Furthermore, the ability to classify individual patient change bridges the gap between traditional efficacy and effectiveness studies (that focus on changes made by groups of patients) and clinical practice (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999).

Detecting Potential Treatment Failure

A core element of outcome management systems is detecting potential treatment failure. To improve the outcomes of patients who are responding poorly to treatment, such patients must be identified before termination from treatment and, ideally, as early as possible in the course of treatment. Many studies have investigated the value of several patient, therapist, patient–therapist interaction, and extratherapeutic variables in predicting outcome, yet very few of the variables explored are consistently highly predictive of outcome. Research utilizing the OQ has indicated that the best predictors of outcome are initial severity of distress (i.e., pretreatment OQ Total Score) and change score after separate sessions early in the course of treatment. In fact, Brown and Lambert (1998) found that pretreatment OQ Total Score and change scores from Sessions 1 to 3 accounted for approximately 40% of the variance in final outcome. After taking these variables into account, all other variables combined (e.g., diagnosis, patient demographics, therapist demographics, therapist theoretical orientation) accounted for less than 1% of the variance in final outcome. In other words, in previous studies using the OQ, the best way to predict outcome was to know how distressed patients were before treatment and whether the changes they made early in the treatment process were positive or negative.

Clearly, potential limitations in the aforementioned research relate to the extent to which these findings may be a function of monosource bias (because the patient was the source of both ratings), common method variance (because the OQ was both the predictor and the outcome criterion), or regression to the mean. However, it is important to note that change scores on the OQ have been found to correlate highly with change scores noted on other measures that are frequently used to assess outcome (Beckstead et al., 2003).

Given research on the variables most predictive of outcome, an empirically derived signal-alarm system was developed to alert clinicians to potential treatment failures. This system plots a statistically generated expected recovery curve for different levels of pretreatment distress on the OQ and uses this plot as a basis for identifying patients who are not making expected treatment gains and, therefore, are at risk for a poor outcome. The accuracy of this signal-alarm system has been evaluated in a number of empirical investigations (Lambert, Whipple, Bishop, et al., 2002; Lutz et al., 2006; Percevic, Lambert, & Kordy, 2006; Spielmans, Masters, & Lambert, 2006). An extensive discussion of the results of these studies is beyond the scope of this chapter. However, it is important to note that the signal-alarm system is highly sensitive in that it is able to

predict accurately a poor outcome in 88% of cases that actually end with a negative outcome (as measured by the OQ). It is also far superior to clinical judgment in its ability to identify patients who are at risk for a negative treatment outcome (Hannan et al., 2005).

The Provision of Feedback to Therapists and Patients

The feedback system has been used as an intervention for preventing deterioration and enhancing positive outcomes in patients in that it alerts clinicians to potential treatment failures and allows them to modify their treatment approach in an attempt to improve the outcomes of patients experiencing a poor response to treatment who are predicted to have a poor outcome. Once a patient takes the OQ, commences treatment, and completes a session of treatment, the signal-alarm system can be implemented. When the OQ is electronically administered, it generates immediate feedback regarding the patient's progress. The feedback to therapists consists of several components, among which are a progress graph that includes all the patient's OQ Total Scores from pretreatment to the current session and a color-coded message (white, green, yellow, or red) that indicates the status of patient progress. The specific language of the feedback messages varies not only as a function of patient progress but also as a function of the session at which the feedback is provided (i.e., a red message at Session 2 is not as urgent as a red message at Session 20). A summary of each feedback message follows:

- White message: "The patient is functioning in the normal range. Consider termination."
- Green message: "The rate of change the patient is making is in the adequate range. No change in the treatment plan is recommended."
- Yellow message: "The rate of change the patient is making is less than adequate. Recommendations: Consider altering the treatment plan by intensifying treatment, shifting intervention strategies, and monitoring progress especially carefully. This patient may end up with no significant benefit from therapy."
- Red message: "The patient is not making the expected level of progress. Chances are he/she may drop out of treatment prematurely or have a negative treatment outcome. Steps should be taken to carefully review this case and decide on a new course of action such as referral for medication or intensification of treatment. The treatment plan should be reconsidered. Consideration should also be given to presenting this patient at case conference. The patient's readiness for change may need to be reassessed."

In addition to providing feedback regarding patient progress to therapists, feedback can also be provided directly to patients (Harmon et al., 2007; Hawkins et al., 2004). Patient feedback graphs and messages (i.e., white, green, yellow/red) that correspond to the aforementioned therapist feedback messages have been developed in an effort to inform patients directly of their progress in treatment (in relation to similar patients) and enhance collaboration between the patient and the therapist, practices that are known to be related to positive outcome in many domains of health care. An example feedback message (without the corresponding graph) that may be provided between Sessions 2 and 4 follows:

> Yellow/Red message: "Please note that the information presented below is based on your responses to the questionnaire that you complete prior to each therapy session." It appears that you have not experienced a reduced level of distress. Because you may not be experiencing the expected rate of progress, it is possible that you have even considered terminating treatment, believing that therapy may not be helpful for you. Although you have yet to experience much relief from therapy, it is still early in treatment and there is the potential for future improvement. However, we **urge** you to openly discuss any concerns that you may be having about therapy with your therapist because there are strategies that can be used to help you receive the most out of your therapy. It may also require your willingness to complete additional questionnaires that may

shed light about why you are not experiencing the expected rate of progress.

The administration of the OQ (whether via paper-and-pencil or computerized), scoring, application of the signal-alarm system, and generation of feedback reports (for therapists and/or patients) can be processed in an integrative and almost instantaneous manner using software called OQ-Analyst (administration of the measure and generation of the feedback report takes a total time of approximately 5–7 minutes). Figure 13.5 depicts a screenshot of a therapist feedback report generated by the OQ-Analyst software. This feedback report illustrates the progress of a patient from intake to Session 8. At Session 8, the patient's degree of deterioration (i.e., an increase of 21 points, from 85 at pretreatment to 106 at Session 8) prompted a Red feedback message to the therapist. The feedback report allows the therapist to view all previous OQ scores and associated feedback messages (e.g., this therapist first received a Red feedback message at Session 2). At every session, the therapist is able to look below the graph and read the message that accompanies the feedback. Feedback messages vary depending on the size of the deviation from expected treatment response (the dark sloping line) and the session of therapy at which this deviation occurs. Patient scores are also displayed in relation to the horizontal line at a score of 63, which, as previously mentioned, represents the cutoff score between patient and nonpatient populations on the OQ. The feedback report also provides information about the patient's answer to five critical items as well as other information (e.g., whether the patient's change at the current session meets clinical significance criteria for recovery, improvement, no change, or deterioration) that may be helpful to a therapist working with such a patient.

Resources for Working With Nonresponding and Deteriorating Patients

Over the past 25 years, methodologies have been used in medical research and practice to manage clinical interventions in areas such as drug dosage, diagnosis, and preventive care. These interventions are often used in a stepwise approach that assists physicians in clinical decision making and provides recommendations to improve the quality of patient health care (Hunt, Haynes, Hanna, & Smith, 1998). Similarly, a set of clinical support tools (CSTs) was developed and integrated into the existing psychotherapy quality management system in an attempt to augment the feedback provided to therapists and further improve outcomes of nonresponding and deteriorating patients (Lambert, Whipple, et al., 2004). As such, CSTs are intended to be utilized by therapists when a patient is predicted to have a poor outcome (i.e., when a therapist receives a Red or Yellow warning message, indicating that the patient is not responding or is deteriorating in treatment).

The CSTs are composed of a problem-solving decision tree designed to direct therapists' attention systematically to certain factors that have been shown to be consistently related to patient outcome in the empirical literature, such as the therapeutic alliance, social support, readiness to change, diagnostic formulation, and need for psychiatric referral. The Assessment for Signal Clients (Lambert et al., 2008), a 40-item measure aimed at assisting therapists in the assessment of constructs known to be related to outcome (e.g., the quality of the therapeutic alliance, patient motivation for change, patient perceptions of social support, patient stressful life events) is also included. Furthermore, the CSTs provide specific intervention strategies that could be used by therapists if problems were detected in the aforementioned domains.

EFFECT OF FEEDBACK ON PATIENT OUTCOME

Five controlled studies have been published that examine the effects of providing patient progress feedback to therapists and patients (Harmon et al., 2007; Hawkins et al., 2004; Lambert, Whipple, Smart, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Whipple et al., 2003). Each of the studies required about 1 year of daily data collection and evaluated the effects of providing feedback about a patient's improvement through the use of progress graphs and warnings about patients who were not demonstrating expected treatment responses (signal-alarm cases). The primary question in each of these

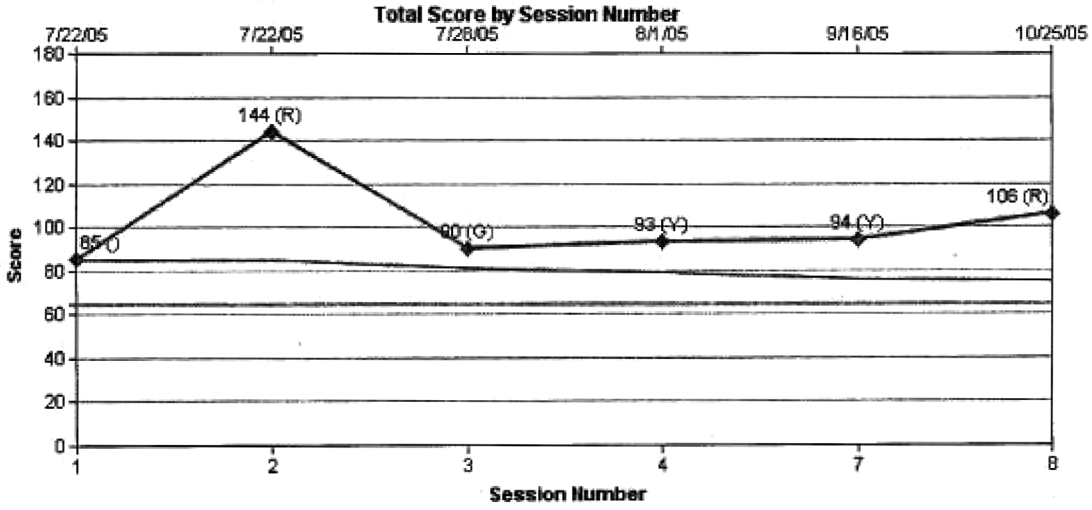
Name: news, brad **ID:**
Session Date: 10/25/2005 **Session:** 8
Clinician: lambert, m **Clinic:** Clinic A
Diagnosis: Unknown Diagnosis
Algorithm: Empirical

Alert Status: **Red**
Most Recent Score: 106
Initial Score: 85
Change From Initial: Reliably Worse
Current Distress Level: High

Most Recent Critical Item Status:

8.	**Suicide** - I have thoughts of ending my life.	**Frequently**
11.	**Substance Abuse** - After heavy drinking, I need a drink the next morning to get going.	**Frequently**
26.	**Substance Abuse** - I feel annoyed by people who criticize my drinking.	**Frequently**
32.	**Substance Abuse** - I have trouble at work/school because of drinking or drug use.	**Sometimes**
44.	**Work Violence** - I feel angry enough at work/school to do something I might regret.	**Sometimes**

Subscales	Current	Outpat. Norm	Comm. Norm
Symptom Distress:	61	49	25
Interpersonal Relations:	26	20	10
Social Role:	19	14	10
Total:	**106**	**83**	**45**



Graph Label Legend:
(R) = **Red**: High chance of negative outcome (Y) = **Yellow**: Some chance of negative outcome
(G) = **Green**: Making expected progress (W) = **White**: Functioning in normal range

Feedback Message:
The patient is deviating from the expected response to treatment. They are not on track to realize substantial benefit from treatment. Chances are they may drop out of treatment prematurely or have a negative treatment outcome. Steps should be taken to carefully review this case and identify reasons for poor progress. It is recommended that you be alert to the possible need to improve the therapeutic alliance, reconsider the client's readiness for change and the need to renegotiate the therapeutic contract, intervene to strengthen social supports, or possibly alter your treatment plan by intensifying treatment, shifting intervention strategies, or decide upon a new course of action, such as referral for medication. Continuous monitoring of future progress is highly recommended.

FIGURE 13.5. A sample OQ analyst feedback report for the therapist.

studies was: Does formal feedback to therapists (and in one study, patients) on patient progress improve psychotherapy outcomes? The prediction in each of these studies was that patients identified as signal-alarm cases (those predicted to have a poor final treatment response) whose therapist received feedback would show better outcomes than similar patients whose therapists did not receive feedback.

The five studies shared many things in common: (a) Each included consecutive cases seen in routine care regardless of patient diagnosis or comorbid conditions (rather than being disorder specific); (b) random assignment of patients to experimental (feedback) and treatment-as-usual (TAU) conditions (no feedback) was made in all but one of the studies; (c) psychotherapists provided a variety of theoretically guided treatments, with most adhering to cognitive–behavioral and eclectic orientations and fewer representing psychodynamic and experiential orientations; (d) a variety of clinicians were involved—approximately 50% of patients were seen by postgraduate therapists, and approximately 50% were seen by graduate students; (e) therapists saw both experimental (feedback) and no-feedback cases, thus limiting the likelihood that outcome differences between conditions could be due to therapist effects; (f) the outcome measure as well as the methodology rules and standards for identifying signal-alarm patients (failing cases) remained constant; (g) the duration of therapy (dosage) was determined by patient and therapist rather than by research design or arbitrary insurance limits; and (h) patient characteristics such as gender, age, and ethnicity were generally similar across four of the studies and came from the same university counseling center, whereas the fifth sample (Hawkins et al., 2004) was older, more disturbed, and treated in a hospital-based outpatient clinic.

A notable difference in the studies was that two of the studies (Harmon et al., 2007; Whipple et al., 2003) included a second experimental condition that was intended to strengthen the feedback intervention by encouraging therapists to use the CSTs (i.e., problem-solving decision tree, additional measures and cutoffs, and suggestions for alternative clinical interventions) with signal-alarm cases. Two of the studies (Harmon et al., 2007; Hawkins et al., 2004) also included two experimental conditions aimed at comparing TAU with feedback to therapists, and feedback to both therapists and patients.

Results from the combined studies are presented graphically in Figure 13.6. As shown, patients identified as not responding or deteriorating (collectively referred to as "not-on-track [NOT]") had a different outcome course depending on assignment to the

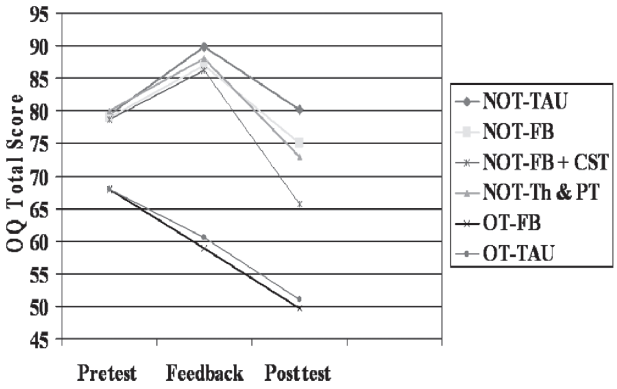


FIGURE 13.6. The effects of providing patient progress information to therapists and patients. Illustrates change from pre- to posttesting of NOT (signal-alarm, not-on-track) and OT (on-track) patients. NOT-TAU = signal-alarm cases whose therapist got no signal or message; NOT-FB = signal-alarm cases whose therapist got a red or yellow signal, indicating they were at risk for treatment failure; OT-TAU = patients who were making satisfactory progress and whose therapist never received any information about their progress; OT-FB = patients whose therapist got a green or white signal and message and who were predicted to have a positive outcome; NOT-FB + CST = signal-alarm cases whose therapist got feedback and used the CST; NOT-Th &PT Fb = signal-alarm patients who entered treatment and were assigned to receive feedback and whose therapist also received feedback; Pretest = average patient scores on the OQ at intake; Feedback = average patient score on the OQ at the point at which a patient qualified for a yellow or red message (the time of warning varied across patients); Posttest = average patient OQ score at the session they terminated treatment (number of sessions until termination occurred varied). OQ Total Score = Outcome Questionnaire Total Score; TAU = treatment as usual; FB = Feedback; CST = Clinical Support Tools; Th = Therapist; PT = Patient.

no-feedback or feedback treatment conditions. Up to the point that these signal-alarm cases were first signaled (or, in the case of the no-feedback treatment, could have been signaled), the graph illustrates an average worsening of around 10 points (about 0.5 standard deviations on the OQ). From the point of the signal-alarm, all the experimental (feedback) groups improved, whereas the no-feedback control TAU cases improved to an average score near 80 but were, as a group, slightly worse off than when they entered treatment. Also displayed is the outcome for on-track (OT) cases where therapists did get feedback and did not get feedback. As illustrated in Figure 13.6, these

patients made steady progress and left treatment, as a group, well within the ranks of normal functioning. It appeared to make little difference in outcome for feedback (Green or White messages) to have been given.

In the individual studies themselves, the effect sizes (Cohen's *d*) for the difference between various feedback conditions for the NOT patients and TAU controls ranged from a low of .34 (when NOT patients whose therapists received feedback regarding their patients' progress were compared with TAU controls whose therapists received no feedback) to .92 (when NOT patients whose therapists received feedback regarding their patients' progress and used the CSTs to improve outcomes in these patients were compared with TAU controls whose therapists received no feedback). Such effect sizes are surprisingly large when one considers that an average effect for comparative studies (active treatments) typically falls between .00 and .20 (Lambert & Ogles, 2004) and is widely considered important enough to lead to a recommendation of a "best practice." Across the five studies, some inconsistent results have been found. Usually the provision of NOT feedback increases the number of sessions that patients attend by about two to three sessions (compared with the NOT no-feedback condition) and decreases sessions for OT cases by two thirds of a session (compared with the OT no-feedback condition), but this result was not observed in the Hawkins et al. (2004) study where number of sessions attended by both NOT groups was equal and number of sessions attended by both OT groups was also equal. In about half the studies, feedback to OT cases improved outcomes despite reducing treatment length. Direct feedback to patients in the form of a written message improved outcomes dramatically in the Hawkins et al. (2004) study but had no effect in the Harmon et al. (2007) replication.

Table 13.1 presents a classification of signal-alarm patients based on their final treatment status at termination. As shown, 20% of the signal-alarm cases seen by therapists who received no feedback showed a negative treatment outcome at termination. In contrast, when therapists received feedback that identified their patient as NOT, only 15% deteriorated; 12% deteriorated when both client and therapist received feedback, and 8% deteriorated when progress feedback and clinical support tools were employed. The rates for signal-alarm cases (NOT) showing clinically significant or reliable change were also markedly different, with the highest rates of improvement in the therapist feedback + CST condition (45%), compared with 22% in the TAU condition. These data suggest that the improved outcomes for patients in the experimental conditions are not only statistically significant but possess considerable clinical meaning for the individual patient.

SUMMARY AND CONCLUSION

Psychological assessment of treatment is a multifaceted process, the success of which is contingent on

TABLE 13.1

Percentage of Not-on-Track (Signal-Alarm) Cases Meeting Criteria for Clinically Significant Change at Termination Summed Across Five Studies

	n (%)			
Outcome classification	**TAU**	**T-Fb**	**T-Fb + CST**	**T/C-Fb**
Deteriorated[a]	64 (20)	90 (15)	12 (8)	19 (12)
No change	184 (58)	316 (53)	73 (47)	71 (46)
Reliable or clinically significant change[b]	70 (22)	196 (33)	169 (45)	57 (37)

Note. TAU = Patients who were not on track and whose therapist was not given feedback; T-Fb = patients who were not on track and whose therapist received feedback; T-Fb + CST = patients who were not on track and whose therapist received feedback and used clinical support tools; T/C-Fb = therapist feedback plus written direct feedback to clients.

[a]Worsened by at least 14 points on the OQ from pretreatment to posttreatment. [b]Improved by at least 14 points on the OQ or improved and passed the cutoff between dysfunctional and functional populations.

several factors. The selection of an appropriate outcome measure for assessing patient progress throughout the course of treatment and at termination is essential to the accurate assessment of treatment effects. Although there are many factors to consider when selecting an outcome measure, the value of any outcome measure is largely related to its ability to be sensitive to the changes patients make after interventions. The use of psychological measures that are inappropriate for assessing patient outcome may lead to inaccurate conclusions regarding the effectiveness of treatment and have significant adverse consequences for patients, therapists, and other important stakeholders in the treatment process.

Once an appropriate outcome measure is selected, a methodology for measuring, monitoring, and managing outcomes can be developed. Defining a positive and negative outcome, developing an empirically based methodology by which to detect patients who are not responding to or deteriorating during treatment as early as possible in the treatment process, regularly utilizing outcome data by means of feedback to therapists and patients, and providing resources to therapists working with nonresponding or deteriorating patients are all central features of any psychotherapy quality management system.

The psychotherapy quality management system described in this chapter was specifically designed with attention to the aforementioned factors and with the explicit intent of monitoring the progress of all patients and improving treatment response for those who are predicted to have a poor outcome. The provision of real-time feedback to therapists and patients in the context of this psychotherapy quality management system has been studied in five clinical trials. Given the large sample sizes of the five clinical trials, and a combined overall sample size of more than 4,000 patients, the present findings seem compelling and suggest that the provision of feedback to therapists in cases that are at risk for treatment failure should be considered an evidence-based practice in psychology ("Evidence-Based Practice in Psychology," 2006; Lambert et al., 2003). It is hoped that the concepts discussed in this chapter will serve as a model and convince researchers and practitioners that systematically monitoring and managing patient outcomes is essential in improving the effects of treatment.

References

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., . . . McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69,* 184–196. doi:10.1037/0022-006X.69.2.184

Beckstead, D. J., Hatch, A. L., Lambert, M. J., Eggett, D. L., Goates, M. K., & Vermeersch, D. A. (2003). Clinical significance of the Outcome Questionnaire (OQ-45.2). *Behavior Analyst Today, 4*, 79–90.

Berrett, K. M. (1998). *Youth Outcome Questionnaire (Y-OQ): Item sensitivity to change*. Unpublished doctoral dissertation, Department of Psychology, Brigham Young University, Provo, UT.

Brown, G. S., & Lambert, M. J. (1998, June). *Tracking patient progress: Decision making for cases who are not benefiting from psychotherapy*. Paper presented at the annual meeting of the Society for Psychotherapy Research, Snowbird, UT.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & Dahlstrom, M. W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring* (rev. ed.). Minneapolis: University of Minnesota Press.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY: Harper & Row.

DuBois, P. H. (1970). *A history of psychological testing.* Boston, MA: Allyn & Bacon.

Evidence-based practice in psychology. (2006). *American Psychologist, 61*, 271–285. doi:10.1037/0003-066X.61.4.271

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D., & Cox, D. (1992). Quality of life measures in health care I: Applications and issues in assessment. *British Medical Journal, 305*, 1074–1077. doi:10.1136/bmj.305.6861.1074

Froyd, J. E., Lambert, M. J., & Froyd, J. D. (1996). A survey and critique of psychotherapy outcome measurement. *Journal of Mental Health, 5*, 11–16.

Guyatt, G. (1988). Measuring health status in chronic airflow limitation. *European Respiratory Journal, 1*, 560–564.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*, 155–163.

Hansen, N. B., Lambert, M. J., & Forman, E. V. (2002). The psychotherapy dose–response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9*, 329–343. doi:10.1093/clipsy.9.3.329

Harmon, S. C., Lambert, M. J., Smart, D. W., Hawkins, E. J., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist/client feedback and clinical support tools. *Psychotherapy Research, 17*, 379–392. doi:10.1080/10503300600702331

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K., & Tuttle, K. (2004). The effects of providing patient progress information to therapists and patients. *Psychotherapy Research, 14*, 308–327. doi:10.1093/ptr/kph027

Hill, C. E., & Lambert, M. J. (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. 84–135). New York, NY: Wiley.

Horowitz, M. J., Milbrath, C., & Stinson, C. H. (1997). Assessing personality disorders. In H. H. Strupp, L. M. Horowitz, & M. J. Lambert (Eds.), *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery* (pp. 401–432). Washington, DC: American Psychological Association. doi:10.1037/10232-015

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist, 41*, 159–164. doi:10.1037/0003-066X.41.2.159

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Efficacy, effectiveness, and client progress. *American Psychologist, 51*, 1059–1064. doi:10.1037/0003-066X.51.10.1059

Hunt, D. L., Haynes, R. B., Hanna, S. E., & Smith, K. (1998). Effects of computer-based clinical decision support systems on physician performance: A systematic review. *JAMA, 280*, 1339–1346. doi:10.1001/jama.280.15.1339

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. doi:10.1037/0022-006X.59.1.12

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. doi:10.1037/0022-006X.67.3.285

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*, 27–36. doi:10.1016/0021-9681(85)90005-0

Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart-Heidelberg model. *Journal of Consulting and Clinical Psychology, 69*, 173–183. doi:10.1037/0022-006X.69.2.173

Kraus, D. R., & Horan, F. P. (1997). Outcomes roadblocks: Problems and solutions. *Behavioral Health Management, 17*, 22–26.

Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In M. J. Lambert, E. R. Christensen, & S. S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 3–32). New York, NY: Wiley.

Lambert, M. J. (2010). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice*. Washington, DC: American Psychological Association.

Lambert, M. J., Bailey, R. J., Kimball, K., Shimokawa, K., Harmon, S. C., & Slade, K. (2008). *CSTs Manual—Brief Version—40*. Unpublished manuscript, Department of Psychology, Brigham Young University, Provo, UT.

Lambert, M. J., Horowitz, L. M., & Strupp, H. H. (1997). Conclusions and recommendations. In H. H. Strupp, L. M. Horowitz, & M. J. Lambert (Eds.), *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery* (pp. 491–502). Washington, DC: American Psychological Association. doi:10.1037/10232-019

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., & Burlingame, G. M. (2004). *Administration and scoring manual for the Outcome Questionnaire-45*. Salt Lake City, UT: OQ Measures.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193). New York, NY: Wiley.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically derived and rationally derived methods for identifying clients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*, 149–164. doi:10.1002/cpp.333

Lambert, M. J., Whipple, J. L., Harmon, C., Shimokawa, K., Slade, K., & Christofferson, C. (2004). *Clinical Support Tools manual*. Provo, UT: Department of Psychology, Brigham Young University.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10*, 288–301. doi:10.1093/clipsy.bpg025

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on client progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*, 49–68. doi:10.1080/713663852

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. K. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9*, 91–103. doi:10.1002/cpp.324

Lipsey, M. W. (1990). *Design sensitivity*. Newbury Park, CA: Sage.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 400–410. doi:10.1037/0022-006X.66.2.400

Lutz, W., Lambert, M. J., Harmon, S. C., Stulz, N., Tschitsaz, A., & Schürch, E. (2006). The probability of treatment success, failure and duration—What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy, 13*, 223–232. doi:10.1002/cpp.496

Meier, S. T. (1997). Nomothetic item selection rules for tests of psychological interventions. *Psychotherapy Research, 7*, 419–427. doi:10.1080/10503309712331332113

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome system. *Journal of Clinical Psychology: In Session, 61*, 199–208.

Milner, J. S., Gold, R. G., Ayoub, C., & Jacewitz, M. M. (1984). Predictive validity of the Child Abuse Potential Inventory. *Journal of Consulting and Clinical Psychology, 52*, 879–884. doi:10.1037/0022-006X.52.5.879

Newman, F. L., & Ciarlo, J. A. (1994). Criteria for selecting psychological instruments for treatment outcome assessments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 98–110). Hillsdale, NJ: Erlbaum.

Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcome assessment*. New York, NY: Wiley.

Percevic, R., Lambert, M. J., & Kordy, H. (2006). What is the predictive value of responses to psychotherapy for its future course? Empirical explorations and consequences for outcome monitoring. *Psychotherapy Research, 16*, 364–373. doi:10.1080/10503300500485524

Pilkonis, P. A. (1997). Measurement issues relevant to personality disorders. In H. H. Strupp, L. M. Horowitz, & M. J. Lambert (Eds.), *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery* (pp. 371–388). Washington, DC: American Psychological Association. doi:10.1037/10232-013

Shea, M. T. (1997). Core battery conference: Assessment of change in personality disorders. In H. H. Strupp, L. M. Horowitz, & M. J. Lambert (Eds.), *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery* (pp. 389–400). Washington, DC: American Psychological Association. doi:10.1037/10232-014

Spielmans, G. I., Masters, K. S., & Lambert, M. J. (2006). A comparison of rational versus empirical methods in prediction of negative psychotherapy outcome. *Clinical Psychology and Psychotherapy, 13*, 202–214. doi:10.1002/cpp.491

Trabin, T. (1995). Making quality and accountability happen in behavioral healthcare. *Behavioral Healthcare Tomorrow, 4*, 5–6.

Tryon, W. W. (1991). *Activity measurement in psychology and medicine*. New York, NY: Plenum Press.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*, 242–261. doi:10.1207/S15327752JPA7402_6

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. doi:10.1037/0022-0167.51.1.38

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Pearson.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59–68. doi:10.1037/0022-0167.50.1.59

CHAPTER 14

PSYCHOLOGICAL ASSESSMENT IN ADULT MENTAL HEALTH SETTINGS

Sandra L. Horn, Joni L. Mihura, and Gregory J. Meyer

A variety of different adult mental health settings exist, including university counseling centers, private clinician practices, community mental health centers, inpatient psychiatric hospitals, day programs, emergency care, psychiatric long-term residential care, forensic and neuropsychological settings. Because other chapters in this handbook focus on specialty settings and types of assessment (e.g., forensic, health, neuropsychological), this chapter presents information about adult psychological assessments in general. When appropriate, the implications of the different settings for adult psychological assessments are addressed.

TESTING VERSUS ASSESSMENT

Although the terms are sometimes used interchangeably, within clinical psychology psychological testing and psychological assessment are distinct clinical endeavors. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association [APA], & National Council on Measurement in Education, 1999) make the distinction as follows:

> A test is an evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process. . . . Assessment is a broader term, commonly referring to a process that integrates test information with information from other sources (e.g., information from the individual's social, educational, employment, or psychological history). (p. 3)

Some also characterize certain models of assessment (e.g., collaborative assessment, therapeutic assessment) as a collaborative process through which therapeutic gains can be made (e.g., Finn, 2007; Fischer, 1994).

The difference between psychological testing and assessment can be demonstrated with a brief example adapted from Meyer et al. (2001): From a psychological testing perspective, the nomothetic meaning associated with a scaled score of 10 on the Arithmetic subtest from the Wechsler Intelligence Scale is that a person possesses average skills in mental calculations. However, in an idiographic assessment, the same score may mean very different things. On the basis of a review of all relevant information, this score may mean that a patient with a recent head injury has had a marked decline in auditory attention span and the capacity to mentally manipulate information. In a patient undergoing cognitive remediation therapy for attentional problems secondary to a head injury, this score may mean that the patient has had a substantial recovery of cognitive functioning. For a third, very intelligent patient with generally elevated subscale scores, this score of 10 may mean that marked symptoms of anxiety and depression are impairing skills in active concentration. It is also possible that, in another patient with similarly high intelligence, a score of 10 might be obtained because the patient was malingering. Meyer et al. continued on to the following conclusion:

DOI: 10.1037/14048-014
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

> Thus, and consistent with Shea's (1985) observation that no clinical question can be answered solely by a test score, many different conditions can lead to an identical score on a particular test. The assessment task is to use test-derived sources of information in combination with historical data, presenting complaints, observations, interview results, and information from third parties to disentangle the competing possibilities (Eyde et al., 1993). The process is far from simple and requires a high degree of skill and sophistication to be implemented properly. (p. 144)

Given the complexities inherent in skilled assessment practice, it is not surprising that psychological assessment was one of the primary domains of training and practice that was a focus of the 2002 *Competencies Conference* (Kaslow et al., 2004), a profession-defining conference cosponsored by more than 35 professional organizations that brought together delegates from a wide range of education, training, credentialing, and practice constituencies. The Psychological Assessment Work Group assembled for that conference subsequently published a set of eight core skill domains that they deemed important for achieving psychological assessment competency (Krishnamurthy et al., 2004, pp. 732–733). These competencies are listed in Exhibit 14.1.

Exhibit 14.1
Summary of Assessment Competencies From the 2002 Competencies Conference

1. A background in the basics of psychometric theory
2. Knowledge of the scientific, theoretical, empirical, and contextual bases of psychological assessment
3. Knowledge, skill, and techniques to assess the cognitive, affective, behavioral, and personality dimensions of human experience with reference to individuals and systems
4. The ability to assess outcomes of treatment/intervention
5. The ability to evaluate critically the multiple roles, contexts, and relationships within which clients and psychologists function, and the reciprocal effect of these roles, contexts, and relationships on assessment activity
6. The ability to establish, maintain, and understand the collaborative professional relationship that provides a context for all psychological activity including psychological assessment
7. An understanding of the relationship between assessment and intervention, assessment as an intervention, and intervention planning
8. Technical assessment skills that include:
   a. Problem and/or goal identification and case conceptualization
   b. Understanding and selection of appropriate assessment methods including both test and nontest data (e.g., suitable strategies, tools, measures, timelines, and targets)
   c. Effective application of the assessment procedures with clients and the various systems in which they function
   d. Systematic data gathering
   e. Integration of information, inference, and analysis
   f. Communication of findings and development of recommendations to address problems and goals
   g. Provision of feedback that is understandable, useful, and responsive to the client, regardless of whether the client is an individual, group, organization, or referral source

Additionally, the work group suggested ways to evaluate competencies, keeping in mind the differing assessment skill demands that occur as a function of setting and educational level (i.e., graduate training programs, internship, and independent practice). At present, there are relatively limited options for documenting competence in general psychological assessment. One option is the American Board of Assessment Psychology (http://www.assessmentpsychologyboard.org), which was organized in 1993 and currently offers diplomate status in psychological assessment for those who have mastered the practice of assessment. However, it was not until 2010 that the APA Council of Representatives voted to add psychological assessment as a recognized area of proficiency. Building on the criteria in Exhibit 14.1, the Society of Personality Assessment (http://www.personality.org) is currently working to define the criteria for proficiency, organize training opportunities, and develop the application procedures to document basic proficiency.

OVERVIEW OF THE NATURE OF PSYCHOLOGICAL ASSESSMENT

Broadly speaking, clinicians engage in informal psychological assessment whenever they work with clients, as they must continually assess what intervention or course of action is appropriate at any point in time. More specifically, however, as described in this chapter, a formal psychological assessment has a circumscribed focus and uses a structured approach to gathering information with the goal of answering one or more specific referral questions. The psychological assessment typically includes oral or written communication of the results (e.g., to the patient or a mental health professional). In contrast to traditional information-gathering models of psychological assessment, contemporary approaches describe the assessment process as an interactive, collaborative, and clinically beneficial two-way joint venture with the patient rather than a one-way attempt to extract information from the patient (e.g., Finn, 2007; Fischer, 1994).

A psychological assessment should be informed by more than one method or source of information. The clinical interview is the most common component of a psychological assessment (Norcross, Karpiak, & Santoro, 2005). In a typical clinical interview with an adult, there are two main sources of information: the patient's self-report and the clinician's behavioral observations.[1] (Interested readers should consult Chapter 11, this volume, for more information concerning the assessment of personality and psychopathology with self-report measures. In addition, portions of Chapter 1, this volume, on clinical and counseling testing, and Chapter 3, this volume, on communicating assessment results, address behavioral observations.) Although a clinical interview is often sufficient to make many of the clinical judgments that are necessary in applied practice (e.g., diagnosis in accordance with the *Diagnostic and Statistical Manual of Mental Disorders* [*DSM*], choice of treatment), additional assessment methods may be needed when the patient's self-reported information is incomplete or insufficient on its own or when it conflicts with other sources of information (e.g., behavioral observations, external sources of information). These additional assessment methods can include interviews with collateral informants, such as a spouse or another mental health professional, or standardized assessment methods such as self-report inventories, observer-rating scales, and performance-based tests of cognitive and personality characteristics.

On the basis of surveys of clinical psychologists, neuropsychologists, and forensic psychologists, the most commonly used broadband tests have been (a) the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV) and the Wechsler Memory Scale—Fourth Edition (WMS) as performance-based cognitive tests (Wechsler, 2008, 2009b); the Minnesota Multiphasic Personality Inventory, original and revised versions (MMPI and MMPI–2, respectively; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Graham et al., 2001; Hathaway & McKinley, 1942) as self-report personality tests; and the Rorschach Inkblot Test (1921/1942) and Thematic Apperception Test (TAT; Murray, 1943) as performance-based personality tests (see Camara, Nathan, & Puente, 2000; Lees-Haley, 1992; Rabin, Barr, & Burton, 2005). However, brief symptom- and disorder-focused tests are also often included in psychological assessments, such as the self-report-based Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996) or Beck Anxiety Inventory (BAI; Beck & Steer, 1993), the observer-rated Conners' Adult ADHD Rating Scales—Observer Form (Conners, Erhardt, & Sparrow, 1999), and the performance-based Conners' Continuous Performance Test—2 (Conners & MHS Staff, 2004). These brief symptom-focused tests are also commonly used to assess treatment progress or outcome.

GENERAL CONSIDERATIONS WHEN SELECTING AND USING ASSESSMENT METHODS

Targeting the Referral Question and Related Psychological Constructs

To appropriately select tests to be used during an assessment, the reason for referral needs to be clear

[1]Both of these sources of data generally inform semistructured interviews. However, fully structured interviews seek to minimize clinical judgment so generally only obtain information based on the patient's self-report.

and agreed upon by the relevant parties (e.g., referral source, client, and guardian when client cannot give informed consent because of competency issues). Once the referral question(s) has been outlined, the clinician can consider which tests may be helpful in the assessment. The clinician begins by identifying the psychological construct(s) to be assessed (e.g., intelligence, visual–spatial processing, depression, psychosis, borderline personality disorder) for which they will choose one or more tests to supplement the clinical interview. The clinician should choose tests that target the specific referral questions by first formulating hypotheses that the tests can "answer." As part of this process, the clinician should also recognize that the method used to assess the construct is a crucial part of its definition (e.g., Kagan, 1988), and typically should strive to include multiple methods (e.g., depression as assessed by self-report, behavioral observation, and the thematic content of narratives).

Contextual Considerations

In choosing assessment methods, the examiner should consider client-, examiner-, and situation-specific factors that might lead to biased or inaccurate scores. Many of these considerations are practical in nature. For example, is the reading level required for the test within the respondent's skill level? Does the client have any visual/hearing/motor limitations that would preclude him or her from performing in a standardized fashion? Could test administration in the client's secondary versus primary language alter the interpretation of items or final scores? More psychologically complex client-specific factors also must be addressed. For example, how appropriate is the length of the test (or battery) for the individual? Are there factors in the setting (e.g., distractions) or client's behavior (e.g., irritation, lethargy, inattentiveness), including their interaction with the examiner (e.g., negative reactions, overly dependent, aggressiveness) that may compromise the validity of this specific test administration? Often, successful administration addressing these kinds of factors can be achieved with only minor adjustments. Nevertheless, the test manual should be examined for recommendations and cautions, and the adjustments should be taken into consideration when interpreting test scores and also described in the assessment report.

In addition to client-specific factors, valid test administration depends on examiner-specific factors. Is the examiner adept at the administration, scoring, and interpretation of the test (or battery)? The test also must be appropriate for the particular setting. For example, because standardized score results on the Millon Clinical Multiaxial Inventory—III (MCMI–III; Millon, Davis, & Millon, 1997) are based on base rates in clinical samples, using the test in a nonclinical setting (e.g., child custody, occupational) will increase the false-positive rate. Finally, the testing environment can also affect test scores (e.g., light quality, noise, test supplies), and every effort should be made to ensure a quiet, well-lit, and otherwise conducive testing environment.

Standardization

The manner in which the test was standardized can have implications for test scores and their interpretation. The examiner should have enough familiarity with psychometric principles to be able to determine whether a test is appropriate for use with a particular client and referral question. There are two aspects of standardization. One relates to standardized test administration. The key consideration here is whether the test instructions and the clinician's training allow for the test to be administered and scored in the same manner that it was for the normative standardization sample (e.g., seating, instructions to the subject, prompts, clarification rules). The second aspect relates to the normative standardization sample itself. Here the key considerations relate to the generalizability of the construct being assessed and the nature of the reference sample. If the construct being assessed is not consistently affected by demographic factors such as gender, age, region of the country, level of education, and ethnicity, then it is generalizable across these factors, which is the case for many personality characteristics in adults. However, if the construct being assessed is sensitive to these background qualities, then it is not generalizable across them, as is the case for many cognitive ability characteristics. For generalizable constructs, establishing normative samples is fairly straightforward, as a reasonably

large and reasonably diverse sample will provide adequate normative data even if the normative sample is not strictly identical to the population at large on demographic background factors. However, for constructs that are sensitive to background factors, one needs to decide what type of reference sample is most important as a point of comparison: the heterogeneous standard of "people in general" or the homogeneous standard of "people most like the client." For the former, which is most common, one would want to ensure that the normative sample is regionally or nationally representative (although globally representative norms are increasingly in use). For the latter, one would want to ensure that the normative sample matches the client on the key demographic factors of interest.

For all normative data, however, a relevant question concerns an adequate size for the normative sample. A statistical answer to this question would consider the standard error of the mean (*SEM*), which indicates on average how far a sample mean is from the true population mean. The formula is: $SEM = SD / \sqrt{N}$, and with 95% certainty, the true population mean will be in the range of plus or minus approximately 2 *SEM*. Thus, for a test with an *SD* of 15, such as an IQ test, one could be 95% sure that the population mean is within ±3 points of the sample mean when the norms are based on 100 people and within ±1.5 points if $N = 400$.

Reliability

Reliability refers to evidence of a test's consistency, which can be defined in various ways. *Test–retest reliability* represents the consistency of test scores over time, which is most relevant to traitlike constructs expected to be relatively stable. Test–retest reliability coefficients can be confounded by factors such as practice effects or memory effects. *Alternate-forms reliability* demonstrates test consistency over item content in that two parallel forms of the test are constructed and scores on the two forms are correlated. *Split-half reliability* indicates consistency over content (similar to alternate forms reliability), but it is assessed within a test—typically by correlating scores from the odd versus even items. Split-half reliability estimates may be appropriate when the test can be split into theoretically equivalent halves. However, not all tests can be split in such a way (e.g., the Rorschach test and the TAT), and it is not appropriate for speed tests (easy items but time limited)—only power tests (where items vary in difficulty and there are generally no time limits).

Internal-consistency reliability is typically measured with coefficient alpha (α) and indicates an average of all possible split-half reliability coefficients. Internal consistency should not be expected when the construct of interest (and the test items) should be heterogeneous (i.e., complex or multifaceted) and for speed tests. Because α is a function both of item consistency and the number of items on the test, it is wise to always consider the average interitem correlation when assessing homogeneity of content. Finally, *interrater reliability* indicates the consistency of observer agreement when the test is independently scored by two or more observers. It is important to demonstrate good interrater reliability for tests involving complex coding and scoring (e.g., magnetic resonance imaging, electrocardiography, Rorschach scoring systems, some Wechsler tests).

All in all, the clinician must decide which reliability estimates are appropriate to consider for the test, given the construct the test is intended to measure and the nature of the test. Such decisions require psychometric knowledge (e.g., Cicchetti, 1994; Clark & Watson, 1995; Schmidt & Hunter, 1996; Schmidt, Le, & Ilies, 2003; Streiner, 2003a, 2003b). Reliability must also be examined to determine whether the coefficients of interest are sufficiently high to warrant clinical decisions. However, although guidelines exist for what constitutes acceptable levels of internal consistency or interrater reliability (Cicchetti, 1994; Clark & Watson, 1995; Streiner, 2003a, 2003b), reliability is a complex issue, and there is not consensus in the field on these issues. Streiner (2003b), for example, suggested that Nunnally and Bernstein's (1994) recommendation of a minimum α of .90 is too high and stated that "except for extremely narrowly defined traits (and I can't think of any), αs over .90 most likely indicate unnecessary redundancy rather than a desirable level of internal consistency" (p. 103).

Validity

Validity represents the accuracy with which the intended construct is measured. Assessment clinicians

should have at least a basic knowledge of the various forms that validity evidence may take to make an informed test selection as well as an understanding of other factors that can affect the validity of a specific test's scores. Historically, validity was classified into categories including *content validity* (appropriate scope of item content), *criterion-related validity* (evidence of concurrent or predictive correlates), *factorial validity* (evidence of the fit of the test's internal item structure with its expected structure), and *construct validity* (evidence of the fit of the test's external correlates with convergent [positive or negative] and discriminant [near-zero] expectations; see, e.g., Bechtoldt, 1959; Campbell & Fiske, 1959; Cronbach & Meehl, 1955). In more contemporary literature, all these forms of validity tend to be discussed as manifestations of *construct validity*, indicating the extent to which the composite of evidence suggests the test measures what it intends to measure, although there is also increasing recognition of *incremental validity*, which is evidence showing that a test or method makes a unique contribution to a particular assessment task (e.g., Hunsley & Meyer, 2003). The topic of validity and the appropriate magnitude of validity coefficients are complex subjects and are addressed further in the section The Multimethod Convergence Problem later in this chapter.

Although a test manual may report excellent validity statistics, validity is context dependent. A test that has high levels of established validity for one type of assessment question may be entirely inappropriate for use in a different assessment. As an example, the Wechsler Memory Scale—Fourth Edition (Wechsler, 2009b) has excellent construct validity when used to assess declarative episodic memory abilities (e.g., remembering events) but not procedural memory abilities (e.g., remembering how to ride a bike or tie shoelaces). Additionally, if multiple memory tests are used within a single assessment, the assessor may be using tests that each individually has high construct validity but also low incremental validity when used in combination because they overlap so much with each other.

Finally, factors present during testing that influence the test scores but do not influence the behavior we want to measure compromise the validity of the test. These are sources of error. It is important to differentiate among various types of error. One primary distinction is between random error and systematic error. Random error refers to inconsistent influences on the test scores observed on that particular test occasion, and in response to it, scores are as likely to increase as they are to decrease. The extent to which a scale is affected by random error is readily quantified by the various types of reliability coefficients, of which retest reliability appears to be the most important (see McCrae, Kurtz, Yamagata, & Terracciano, 2011). Systematic error, on the other hand, refers to any irrelevant constructs or influences that systematically affect the observed test scores on that particular test occasion. In response to it, scores will either consistently increase or decrease; they will not do both. These types of errors are very difficult to quantify. However, they may be built into the test itself (e.g., by having skewed item content; by requiring participants to have a certain level of self-awareness for accurate reporting), which compromises its nomothetic validity, or they may be present when assessing a particular person on a particular occasion (e.g., by transient fear of not getting needed treatment prompting a patient to overreport problems on a self-report symptom scale; by chronic oppositionality making a client unwilling to fully engage in a performance task), which compromises its idiographic validity. From this perspective, every observed test score (X) can be thought of as comprising three components that vary in the extent of their contributions: the construct of interest that one hopes to measure (CI), systematic error (*SE*), and random error (RE). Thus, an observed test score can be then written as X = CI + *SE* + RE. The extent to which a test is reliable is determined by CI + *SE* (what is called the "true score" in the true score theory of measurement, although "true" is best understood as "consistent" or "unwavering," not "accurate"; Allen & Yen, 1979). The extent to which a test is valid in practice is determined only by CI. To illustrate the components of the equation, consider the following example using the BDI (see Exhibit 14.2).

The Cross-Test Problem of Different Metrics

When integrating assessment information from various interviews and tests, the clinical picture

Exhibit 14.2
Test Score Components

X	=	CI	+/–	*SE*	+/–	RE
BDI Score	=	Depression as conceptualized by the BDI	+/–	Social desirability Deliberate manipulation Distrust of examiner Physical illness etc.	+/–	Fatigue Memory Carelessness Environmental distractions etc.

can become difficult to piece together because of inconsistencies across test metrics. Although most psychological tests provide guidance for converting raw scores to standardized scores (e.g., *z* scores, *T* scores, or standard scores) to enable meaningful interpretation of test results, most self- and observer-report indices use *T* score conversions ($M = 50$, $SD = 10$), whereas performance-based tests typically use standard scores ($M = 100$, $SD = 15$). Additionally, some tests do not include uniform instructions for converting raw scores into standardized scores (e.g., the Beck Depression Inventory—II [BDI–II], the Paced Auditory Serial Addition Task, the Test of Memory Malingering). Because one cannot compare raw scores across different tests, forms, raters, and scales in a meaningful way, converting to standardized scores is highly recommended. However, it is imperative that assessors note that standardized scores do not always get scored in the same direction across test scales. For some scales, a high standardized score indicates impairment, whereas on other scales, a low standardized score indicates impairment. For this reason (among others), it is important to check the test manual/guidelines before interpreting scores, even if they are standardized scores. Clinicians who regularly perform assessments of a particular type may find it helpful to create and use a standardized form for plotting testing results from across various types of tests. An example of such a form for assessing adult attention-deficit/hyperactivity disorder (ADHD) is provided in Appendix 14.1, and its application is described in the section Communicating Information later in this chapter.

The Multimethod Convergence Problem

A vital consideration when conducting a psychological assessment is appreciating the method used to collect the assessment data. Although the method(s) of data collection can be easily overlooked by inexperienced assessors, in most assessments, it represents an important source of variance. When properly conceptualized, this variance can help paint a fuller and more accurate clinical picture; if ignored, it can lead to confusion due to seeming inconsistencies across sources of data.

Different Assessment Methods

Primary methods used to obtain assessment data are clinical interviews (structured, semistructured, and unstructured), self-report inventories, observer-report inventories, behavioral observations, and performance-based tests and tasks. Self-report and observer-report assessment methods represent what the reporter is consciously aware of and willing to share. Whether the information is reported to an interviewer or on an inventory, self- and other-report rely on the retrieval of information from memory stores and accurate communication of that information from the reporter. Instead of assessing what the client thinks about or is willing to say about himself or herself, performance-based measures (e.g., Wechsler tests; Rorschach Inkblot Test) assess what the person does behaviorally when provided with a structured task. A benefit of standardized performance measures is that they provide the clinician with information about various psychological characteristics independent of the subjective perception of the clinician, an observer, or the client himself or herself. Readers may consult Chapter 10,

this volume, concerning performance-based personality measures.

The behavioral observations made by the clinician during the interview also can be informative, although the clinician should evaluate the generalizability of the interview situation including how their professional role and their interpersonal style might affect the interview dynamic. Finally, case records taken from intake reports, psychotherapy notes, nursing and psychiatric notes, hospitalization reports, and so forth, can provide valuable historical information and a broader sample of behavior (e.g., behavioral observations, family-member-reported concerns, self-reported symptoms, historical diagnoses, and medications). However, an important limitation of these records is that the clinician who obtains the material is not necessarily able to evaluate its accuracy, the conditions under which it was obtained, or its completeness with respect to the client's past treatment history.

Convergence Across Methods

Typically, when attempting to assess the same or a very similar construct, one finds that the convergence across assessment methods is lower than the convergence within a single method (i.e., the correspondence between self-reported intelligence and performance-based assessment of intelligence is lower than the correspondence between two self-report scales of intelligence or two performance tests of intelligence). In a systematic review of the literature of the validity of psychological and medical assessment methods, Meyer et al. (2001) found a wide range of validity coefficients across tests, regardless of test method. Hemphill (2003) summarized Meyer et al.'s findings to show that the middle range of these cross-method validity effect sizes (r) in psychology was .21 to .33. Of note, however, there was generally low-to-moderate agreement between tests that assessed the same or similar constructs but that made use of distinct methods. For example, correlations were moderate between self-rated and parent-rated personality characteristics (r = .33) and between self-rated and peer-rated personality and mood (r = .27). At the lower end of the range of agreement were self-report and cognitive tests of attention (r = .06) and memory (r = .13; also see Beaudoin & Desrichard, 2011, for a more recent meta-analysis).

The relatively low convergence across assessment methods is not a problem specific to clinical psychology. For decades, experimental psychology researchers have recognized that different methods result in different types of psychological data, and they do not expect self-reported attributes to show strong convergence with externally assessed attributes of the same or similar construct (e.g., Dunning, Heath, & Suls, 2004; Nisbett & Wilson, 1977; Wilson & Dunn, 2004). This finding suggests that psychologists can run into a problem when they rely solely on self-reported information to predict behavior. Many researchers have found that personality attributes that are externally assessed (e.g., with observer ratings, behavioral counts, some Rorschach scales) show significantly stronger levels of convergence with each other than they do with self-report methods (e.g., Connelly & Ones, 2010; Kolar, Funder, & Colvin, 1996; Mihura, Meyer, Dumitrascu, & Bombel, 2012; Riggio & Riggio, 2002).

Although there is an abundance of research demonstrating what we have described as the multimethod convergence problem, this information still seems underappreciated by many clinicians conducting psychological assessments. However, clinicians should seek to capitalize on the uniquely different types of information provided by the different methods. As stated by Meyer et al. (2001), "The quality of idiographic assessment can be enhanced by clinicians who integrate the data from multiple methods of assessment," and that "when assessors systematically integrate this information, they are forced to consider questions, symptoms, dynamics, and behaviors from multiple perspectives—simply because everything does not fit together in a neat and uncomplicated package" (p. 150).

The Role and Limits of Clinical Judgment

For any psychological assessment, all client behaviors and test responses need to be interpreted by the clinician. Thus, the clinician has an important obligation to make well-informed interpretations and decisions based on the various sources of data. Research indicates that professional judgment is not as trustworthy as we as a profession would like to believe.

It is important to understand, however, that judgment bias and inaccuracy is a human problem, not something specific to clinicians. That is, "clinical judgment" is not different from "human judgment." Within cognitive psychology, there is a rather large literature that documents common errors in human reasoning (e.g., Tversky & Kahneman, 1974). Although this research has received substantial criticism (cf. Hammond, 1996; Koehler, 1995), and many instances of cognitive bias can be made to disappear (Gigerenzer, 1991, 2008), there is no doubt that human beings are prone to errors of judgment and reasoning. Clinicians are not immune to these problems (Garb, 1998; Meehl, 1954).

The classic work by Meehl (1954) found that clinicians are less accurate than statistical predictions when interpreting history and test data to predict outcomes (prison recidivism, psychotherapy benefit, military placement, college success). Although such research suggests that using statistical prediction formulae in clinical practice is optimal, there are important considerations to place clinical judgment research in context. First, on average, the extent to which statistical decision rules outperform clinical judgment produces small effect sizes ($d = .12$, which is equivalent to 1.2 points on the *T* score metric; see Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). In fact, as Ægisdóttir et al. (2006) noted, "When judgments are made by expert clinicians, the difference between clinical and statistical methods seems to disappear. However . . . nonexperts . . . are consistently outperformed by . . . formulas" (p. 366). Second, generally, the goal of a clinical evaluation is to develop a multifaceted and idiographic clinical description of the individual, not to predict a specific outcome. Formulae do not exist for this purpose. Third, clinicians can also be sensitive to salient current life events that may significantly alter the clinical presentation, whereas an actuarial formula cannot. Finally, and perhaps most important, there simply are not replicated statistical prediction formulae for the vast majority of judgments that have to be made over the course of a typical psychological assessment.

Although clinicians are not more prone to make judgment errors than other professionals, they can err in a number of different ways. First, clinicians can fall prey to confirmatory biases and, thus, just elicit the kind of information that confirms their hypotheses and hunches. Second, when making diagnostic or classification judgments, clinicians may rely on prototypes or exemplars and judge the fit of their patient to these prototypes rather than systematically evaluating their patient on specific diagnostic criteria. Third, some clinicians are prone to overconfidence bias. Fourth, there is hindsight bias, in which people wrongly assume that they could have predicted an event after being told of the eventual outcome. Finally, clinicians can make erroneous judgments or predictions by not considering the relative frequency of the events they are judging; rare, or low base rate, events are harder to accurately predict than more common events.

The assessment clinician can minimize the errors mentioned earlier by using several corrective strategies (see also Borum et al., 1993; or Spengler, Strohmer, Dixon, & Shivy, 1995). Specifically, assessment clinicians should do the following:

1. Learn as much as possible (theoretically and empirically) about the characteristics of the clinical condition they are evaluating.
2. Directly link test indicators and their absence to the target characteristics.
3. Actively challenge their impressions, or seek to disconfirm their hypotheses, by considering test data that may temper or counter their hypotheses.
4. Take into account the relative frequency of the events or conditions they are trying to predict and make predictions sparingly for rare low base rate events.
5. Use empirically validated statistical predictions whenever possible.
6. Anticipate making errors of judgment and be open to corrective feedback.
7. Actively solicit corrective feedback (from clients and referral sources) to maximize the accuracy of test-derived impressions.

PHASES IN CLINICAL ASSESSMENT

Evaluating the Referral Question

The first phase of psychological assessment is to outline with the referral source the purpose of the

assessment. The examiner needs to know the specific question(s) the client and/or outside parties would like the assessment to address in order to conduct a useful assessment that will meet the needs of those involved. At this stage, the assessor must consider whether there may be possible factors driving the assessment request beyond what is explicitly stated by the referral source or client. The examiner also needs to determine whether the referral source understands psychological testing and psychopathology sufficiently to make a detailed and appropriate referral. In community mental health settings, common referral questions include the following examples: "What is this person's diagnosis?" "Should we consider a medication change?" "Can she work?" "Can he benefit from therapy?" It is the assessment clinician's responsibility to clarify vague referral questions and to elicit and clarify the kind of information the client or referral really wants to know. In general, to do so, the clinician asks questions about the circumstances under which the referral question arose and how the assessment information will be used.

Acquiring Knowledge Relating to the Content of the Problem

The second phase also takes place before beginning any testing. The examiner must understand (as completely and in as much detail as possible) the clinical condition that is being assessed. Furthermore, the examiner must understand what test(s) can assess this condition, what the limitations of the tests are, and how the tests may or may not apply for this particular individual. Thus, the examiner needs to consider questions such as: What am I trying to measure? How will this construct manifest itself? How will this construct be distinct from other similar constructs (e.g., Alzheimer's dementia from multi-infarct dementia)? What relevant but imperfect instruments can I use to quantify this construct? Can I circumvent the shortcomings of any single procedure by using multiple procedures to quantify this construct? (If not, how will I temper my conclusions?) What is the reliability and validity of the measures? Are the norms good? Are there characteristics of this client that makes use of these tests inappropriate?

Meeting the Client/Establishing a Collaborative Contract

Assuming that the client did not initiate the referral, the third phase of the assessment typically involves meeting the client and developing a working relationship. Perhaps the most important point is that assessment is not something a clinician does *to* a person; it is something a clinician does *with* a person. This distinction is central in the collaborative assessment models (see Finn, 2007; Fischer, 1994). It is essential for the client and examiner to have an attitude of mutual exploration with the goal of helping the client understand something about himself or herself that is relevant to the purpose of the assessment. The assessor needs to respect the client, think in terms of adaptation rather than pathology (i.e., frame client behaviors in a nondemeaning fashion), and then begin the process of collaborative exploration. Practically, at minimum, the collaboration should include explaining the purpose of the assessment and of the test administration(s), asking the client for his or her thoughts on findings the clinician does not understand or has difficulty reconciling with the rest of the clinical picture, asking the client for their own hypotheses, and assessing any concerns at the outset. To collaborate successfully, it is also recommended that the examiner contend with their own personal issues that might be provoked in an assessment context (e.g., some common ones are obsessionality, saintliness, forced neutrality, need for approval, voyeurism, control, superiority, narcissism).

Data Collection

Data collection should begin after a working relationship is established with the client. A clinical interview with the client and test administration and scoring is part of data collection, but additional data should also be collected. Other sources of information that can be used include interviews from important others (e.g., spouse, siblings, friends), behavioral observations from each assessment session, and chart and historical information (school records, treatment summaries, previous assessment reports, military records, hospital discharge summaries, prescription history, etc.). The data collection phase can also include the use of tests as methods of

intervention (e.g., Finn, 2007; Fischer, 1994). Although some clinicians may use a fixed battery with every patient when the assessments are conducted for research purposes or when there is a limited range of disorders (e.g., posttraumatic stress disorder assessment in a Veterans Affairs setting), clinicians working in a general clinical setting often use a flexible battery approach or a combination of a fixed and flexible battery. In other words, the clinician may start with a set of initial assessment methods and tests that generally relate to the referral question and then add focused measures as the clinical picture becomes clearer to home in on answering the referral question.

Interpreting the Data

At each step of the way, data collection is followed by interpretation and integration of the data. This phase should be undertaken with the goals of addressing the specific overt and covert referral question(s), accurately understanding and describing the client using professional but easily digested language and terminology, making judgments regarding etiology and prognosis, and detailing appropriate and feasible treatment recommendations. To do this task effectively, the clinician needs to be mindful of the reliability and validity for the various pieces of data being considered, which should be interpreted and integrated with the goal of forming a cohesive clinical picture that encompasses as much of the data as possible, while simultaneously guarding against the tendency to focus on evidence that confirms one's preliminary or initial hypotheses. There is nothing worse than conducting a thorough testing-based assessment but then ignoring the data by simply interpreting results in light of what one already believes on the basis of just the clinical history. As part of a collaborative assessment, clients should actively participate in the development and testing of hypotheses.

Regarding the integration of data, Meyer et al. (2001) emphasized that

> clinicians must consider the nature of the information provided by each testing method, the peculiarities associated with the specific way different scales define a construct . . . the motivational and environmental circumstances that were present during the testing . . . test-based conceptualizations must be reconciled with what is known from history, referral information, and observation. Finally, all of this information must be integrated with the clinician's understanding of the complex condition(s) being assessed. (p. 150)

Integrating assessment findings across sources of information is one of the more advanced clinical skills that a psychologist can acquire. For further information, Finn (1996a) and Ganellen (1996) are both excellent resources for guidance about integrating data across methods, with each using the MMPI–2 and the Rorschach as examples of self-report and performance methods of personality assessment.

Communicating Information

The last assessment phase is to communicate with the client and referral source about the conclusions reached from the assessment. This phase includes writing the assessment report. It is common for clinicians to write the report with the referral audience in mind, but a copy of the report should also be written with the client's reading level in mind. In many cases, it is appropriate to include a summary section in the report that is devoted to the client as the primary audience. Within a collaborative assessment model, the clinician's hypotheses should be discussed with the client throughout the assessment and the client should be encouraged to develop and share their own hypotheses (e.g., Finn, 1996b, 2007; Fischer, 1994). The client's hypotheses and reactions to the clinician's hypotheses should be integrated into the report. Most important, the report should be written with respect for the client's struggles of life rather than from a pathologizing perspective. The collaborative assessment literature provides some excellent insights and suggestions regarding summary sessions where feedback is discussed, but testing should not begin unless the assessor is prepared to sit down with the client and tell him or her the conclusions reached from the assessment.

In clinical settings where clinicians may be repeatedly conducting similar types of assessments

(e.g., ADHD assessments), we recommend using a standardized form for recording raw scores, converting to a standard metric, and plotting test results. An example of such a form is included at the end of this chapter (see the ADHD Summary Score Sheet for Adults in Appendix 14.1). Such standardized forms can simplify cross-test comparisons; provide a framework for visualizing the various test results; and encourage a consistency between and within clinicians in the way that information is recorded, processed, and reported. This consolidated visual layout of the assessment information also can be useful when discussing assessment results with clients, as many clients seem to appreciate the visual presentation of information. Clients may also better comprehend test results if data are presented to them both orally and visually. The assessment results can also provide guidance to the assessor about the kind of feedback that will be most helpful for the client to understand and process, including the method of presenting the information (e.g., visual or verbal) and the nature of the information (e.g., level of complexity, potential client biases in understanding the results, likely reactions to the professional providing the feedback).

With regard to report writing, we suggest that clinicians use the familiar hourglass shape that is used in research reports: The broader clinical history and context for the assessment is at the apex, followed by the referral questions and impressions generated from the preceding information. The narrow middle section of the report includes the tests administered and relevant behavioral observations (as a Method section) and the assessment findings (as a Results section). After this section comes the summary impressions and recommendations sections, which extend outward from the test-based findings in broader ways that link the results with life circumstances and anticipated interventions.

Report writing and feedback sessions are complex undertakings that require much more guidance and preparation than can be provided in this chapter. There are a number of excellent sources detailing general integration of data and report writing (e.g., Blais & Smith, 2008; Kvaal, Choca, & Groth-Marnat, 2003; Lichtenberger, Mather, Kaufman, & Kaufman, 2004). In addition, further discussion concerning the communication of test findings may be found in Chapter 3 of this volume.

Overview of Methods and Measures

The sections that follow review some of the most popular components of a psychological assessment. The sections are organized by assessment method, including the clinical interview, behavioral observations, self-report and observer-report inventories, and performance-based tests.[2] The purpose of these sections is to provide an overview. For further information about the individual methods, the reader can refer to other chapters in this volume, the cited test manuals, or other psychological assessment reference resources (e.g., Groth-Marnat, 2009).

Clinical Interview and Behavioral Observations

The clinical interview serves as the context for the psychological assessment. A clinical interview with an adult uses two main assessment methods: self-reported information by the patient to the interviewer and behavioral observations of the patient by the interviewer. Readers may refer to Chapter 7 of this volume for further discussion of clinical interviewing and related considerations.

Different components of clinical interviews can be unstructured, semistructured, or fully structured. The most common structured component of a clinical interview evaluates *DSM* criteria. Common semistructured interviews for adults include the Structured Clinical Interview for *DSM–IV* Axis I (SCID–I; First, Spitzer, Gibbon, & Williams, 1996) and Axis II (SCID–II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997), among others. The standardized, systematic nature of structured interviews helps increase the reliability of the interviewer's diagnostic judgments (e.g., Lobbestael, Leurgans, & Arntz, 2011). Common "unstructured" components of a clinical interview consist of gathering relevant idiographic history and situational factors and developing and maintaining an alliance with the

[2]Historically, the Rorschach and the Thematic Apperception Test have been referred to as "projective" tests. In keeping with contemporary recommendations, we use the term *performance-based* test instead (Meyer & Kurtz, 2006).

patient. In keeping with the goals of the assessment, the content of the interview should be determined by the referral question.

Just as every psychological test has limitations, clinicians must keep in mind the limitations inherent in interviews and behavioral observations. Perhaps the largest concern with interview data is the inaccuracies that occur because of the use of retrospective recall. Clinical interviews require a person to recall not only past mental states, behaviors, and mood but also to screen the information and choose what they consider most relevant and then summarize that information for the clinician. The memory literature shows that recall accuracy tends to suffer when recall intervals are longer (Brown, Rips, & Shevell, 1985) as well as when dates (Friedman, 1993) or subjective states as opposed to objective facts (Brewer, 1988) are the object of recall.

Seemingly random inaccuracies should be anticipated during clinical interviews, but report bias can also temper the information obtained. For example, there is some evidence that depressed patients overestimate not only past negative affect but also past positive affect, although to a lesser degree (Ben-Zeev, Young, & Madsen, 2009). As another example, Stone et al. (1998) observed that real-time momentary coping reports had little alignment with retrospective reports of coping obtained over the past 48 hours. When young adults and their parents retrospectively recalled ADHD symptoms, accuracy of recall was limited and severity of current symptoms influenced reporting of past symptoms (Miller, Newcorn, & Halperin, 2010). Therefore, the assessor should corroborate interview information with other methods of assessment and review the material for consistency and fit with the empirical literature about the condition or diagnosis.

Throughout the assessment process, behavioral observations provide vital information and serve as a context for the information obtained in the interview and on the psychological tests. Given the unstandardized setting in which the opportunity for behavioral observations evolve, the novice assessor is often challenged to understand which behavioral observations are important. Commonly reported behavioral observations include the client's appearance (e.g., neat, disheveled); whether the patient was oriented to person, place, and time; and characteristics like cooperativeness and how reliable the person was as an informant.

However, the nature of the behavioral observations that are important for any one assessment depend on the specific case and the referral question. For example, if the presenting problem was depression yet the patient appeared cheerful, the assessor would want to understand this discrepancy and note these observations in the report. As another example, if a patient were being assessed for ADHD, and during tests of attention and concentration the assessor observed frequent self-critical comments and visual signs of anxiety such as hand wringing and sweating, the examiner might evaluate a differential anxiety diagnosis and note these observations in the report. Behavioral observations that are congruent with the presenting problem (e.g., tearfulness and depression) are also important behavioral observations that could also provide information about the related context (e.g., the topic that elicited the tearfulness).

Self-Report Inventories

Personality Assessment Inventory (PAI). The PAI (Morey, 2007) is a 344-item, Likert-type self-report inventory appropriate for use with individuals 18 years or older. The test is a broadband measure of personality and psychopathology that clinicians can use as part of a full assessment or as screening measure during intakes. The 22 nonoverlapping full scales include four validity, 11 clinical, five treatment consideration, and two interpersonal scales. Ten of the scales contain subscales. The PAI can be administered by computer or by using an item booklet and answer sheet, with both hand scoring and computer scoring available. Because of nonoverlapping scales and a systematic layout of the response forms, hand scoring is fairly efficient. The PAI Software Portfolio has additional options for comparing the client's results to different reference samples.

MMPI–2 and MMPI—Restructured Form (MMPI–2–RF). The original MMPI was published in 1942 (Hathaway & McKinley, 1942) and substantially revised and reintroduced as the MMPI–2 in

1989 (Butcher et al., 1989; Graham et al., 2001). The MMPI–2 is a 567-item, true–false self-report inventory that contains a total of nine validity and 112 clinical scales. In 2008, a restructured MMPI–2 consisting of a 338-item subset was published (MMPI–2–RF; Ben-Porath & Tellegen, 2008). The MMPI–2–RF has a total of eight validity scales and 42 clinical scales. Like the PAI, the MMPI–2, and MMPI–2–RF can be used for clinical screenings and as part of a larger assessment battery in more complex cognitive and personality assessments. The MMPI–2 and MMPI–2–RF can be administered by computer or by item booklet and answer sheet. Although there are hand-scoring forms available for these tests, we highly recommend that computer scoring software be used as hand scoring is a time-consuming, complex process that can result in errors.

NEO Personality Inventory—Third Edition—Form S (NEO PI–3–S). The NEO PI–3–S is a 240-item Likert-type self-report inventory (Costa & McCrae, 2010) that can aid the clinician in the assessment of normal personality traits using a dimensional model in which both high and low scores are interpretable. The NEO PI–3 assesses the five major personality factors: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Each of the five factors consists of six facet scores. A shortened version of the test consists of 60 items: the NEO Five-Factor Inventory—3 (NEO FFI–3). Scoring can be completed by hand or computer program.

Inventory of Interpersonal Problems—Short Circumplex Scales (IIP–SC). The IIP–SC (Soldz, Budman, Demby, & Merry, 1995) is a 32-item short form derived from the original IIP (Horowitz, Rosenberg, Bauer, Ureño, & Villaseñor, 1988). Four items load on to each of the eight circumplex octants: Domineering, Vindictive, Cold, Socially Avoidant, Nonassertive, Exploitable, Overly Nurturant, and Intrusive. Brief interpersonal measures such as the IIP–SC allow the clinician to quickly screen for various types of interpersonal problems and related distress and can also be useful in delineating the degree to which a client's distress might result from interpersonal (as opposed to non-interpersonal) factors. When interpersonal concerns are causing a client to experience distress, the IIP–SC can be used to track treatment progress.

BDI–II. The BDI–II (Beck, Steer, & Brown, 1996) is a self-report inventory of the presence and intensity of depression symptoms. It is a helpful measure for depression screening and depression symptom monitoring over time. The inventory is quite brief at 21 statement clusters, requiring 5 to 10 minutes for completion. The BDI–II is hand scored, and the total raw score can be compared to suggested cutpoints.

BAI. The BAI (Beck & Steer, 1993) is a self-report inventory used in the assessment of anxiety symptom severity. As with the BDI–II, the BAI is not a diagnostic measure but is appropriately used for symptom screening and monitoring across time. At 21 brief items, the BAI typically requires no more than 5 to 10 minutes to complete. The BAI is hand scored, and the results can be compared with suggested cutpoints.

Conners' Adult ADHD Rating Scales—Self-Report (CAARS–S). The CAARS–S (Conners et al., 1999) is a self-report inventory that assesses behaviors and problems associated with adult ADHD. The self-report CAARS is available in three versions: The 66-item Long Version, the 26-item Short Version, and the 30-item Screening Version. The Long Version contains four factor-derived subscales (Inattention/Memory Problems, Hyperactivity/Restlessness, Impulsivity/Emotional Lability, and Problems with Self-Concept), three ADHD symptom subscales in accordance with the fourth edition of the *DSM* (*DSM–IV;* Inattentive Symptoms, Hyperactive–Impulsive Symptoms, and Total ADHD Symptoms), an ADHD Index, and an Inconsistency Index. The Long Version is recommended when used as part of a battery for making initial diagnostic and treatment decisions and as a baseline measure. CAARS inventories can be scored by hand or computer scoring program.

Observer-Rating Inventories

NEO Personality Inventory—Revised (NEO PI–3–R). This is a parallel observer-report version of the self-report NEO PI–3–S, with items written in the third person. The responder is typically a friend, spouse, or professional, with separate forms for male and female clients.

CAARS—Observer (CAARS–O). The CAARS–O is an observer-report inventory that parallels the self-report CAARS–S, including the three different versions. The CAARS–O can be used as part of a larger test battery for assessment of problems associated with adult ADHD. It is recommended that at least one CAARS–O be used in adult ADHD assessments and that the observer(s) be quite familiar with the person being assessed.

Performance-Based Tasks

WAIS–IV. The WAIS–IV (Wechsler, 2008) is a performance-based test that can be used to assess cognitive ability in adults (ages 16 and up). The test is commonly used as part of specific neuropsychological assessments, ADHD evaluations, learning disorder assessments, and to provide specific information about a client's cognitive strengths and weaknesses. The WAIS–IV is individually administered, which typically requires 1.5 to 2 hours. The test has 15 subtests (5 are optional) that are used to evaluate specific cognitive abilities. The Full-Scale IQ score (FSIQ) is calculated using scores from the Verbal Comprehension Index, Working Memory Index, Perceptual Reasoning Index, and Processing Speed Index. The General Ability Index is an optional index of overall ability that can substitute for the FSIQ and is computed from Verbal Comprehension Index and Perceptual Reasoning Index subtests.

Wechsler Memory Scale—Fourth Edition (WMS–IV). The WMS–IV (Wechsler, 2009b) is a performance-based test of adult memory function and processes. It measures primarily the ability to learn and retrieve specific visual or auditory material initially encountered in the context of the test session, either immediate or delayed (i.e., it largely measures declarative episodic memory). The WMS–IV is commonly used as part of a larger cognitively focused test battery. Administration time varies according to a number of examiner and examinee factors, including how many subtests are relevant for the assessment. There are two administration versions of the WMS–IV: the Adult Battery for ages 16 to 69 and the shorter Older Adult Battery for ages 65 to 90.

Wechsler Individual Achievement Test—Third Edition (WIAT–III). The WIAT–III (Wechsler, 2009a) is a performance-based test of academic achievement that requires approximately 1.5 to 2 hours to administer. The WIAT–III includes 16 subtests that are traditionally organized into seven domain scores assessing oral language, total reading, basic reading, reading comprehension and fluency, written expression, mathematics, and math fluency. However, a factor analysis of the scale intercorrelations reported in the test manual suggests that these domains are not homogenous clusters and that it may be best to consider the subtests as indicators of one overall level of academic achievement (Meyer, 2010).

Conners' Continuous Performance Test—2 (CPT–2). The CPT–2 (Conners & MHS Staff, 2004) is a visual-motor measure of attention, concentration, resistance to monotony, and reaction speed and accuracy. The test consists of a series of letters presented successively on a computer screen, at different time intervals and in a seemingly random order. The test taker is asked to press the space bar on the keyboard every time he or she sees a letter other than the letter "X." Administration time is about 15 minutes. The computer-scored CPT–2 provides 15 scores, including validity checks, although the primary variables are the number of correct responses (hits), the number of times that the space bar was not hit when non-X letters were presented (errors of omission), and the number of times that the space bar was hit when an X was presented (errors of commission). It is not unusual for one or two of the 15 scores to be atypical in the profile of a normally functioning adult. The pattern and elevation level of scores is of great importance, and the test manual should be used for guidance when interpreting CPT–2 profiles.

Paced Auditory Serial Addition Task (PASAT). The PASAT (e.g., Levin et al., 1987) is an auditory measure of attention, concentration, working memory, and speed of information processing; numerous versions of the test are now available, including several computer-administered versions (e.g., Wingenfeld, Holdwick, Davis, & Hunter, 1999). During the testing (with Levin et al. version), a recording is played for the test taker in which four series of 50 numbers (trials) are presented at increasing speed.

Test of Memory Malingering (TOMM). The TOMM (Tombaugh, 1996) provides a cost-effective rough screening for malingering during cognitive assessment. The TOMM is introduced as a test of the person's ability to learn and to remember pictures. The examiner presents 50 simple pictures one at a time followed by a recognition trial in which the test-taker has to identify which of two pictures was presented before. The test is then repeated for a second trial. A third retention-only trial is optional. If malingering is suspected, additional malingering testing and evaluation is strongly advised.

The Rorschach Inkblot Test. The Rorschach Inkblot Test (Rorschach, 1921/1942) is a performance-based measure in which a person is presented with a standard series of 10 inkblots and asked to answer the question "What might this be?" The Rorschach cards contain complex structural elements (e.g., form, color, shading) and allow test takers wide latitude to perceive and organize the stimulus features. As a result, it provides an in vivo sample of behavior obtained under standardized conditions that can be coded along many dimensions (e.g., perceptual, logical, organizational, thematic). The popularity of the Rorschach test in clinical settings despite recurrent psychometric challenges (e.g., Lilienfeld, Wood, & Garb, 2000) is likely due to its ability to provide a method of gathering behavioral information about an individual that cannot be obtained using other popular assessment methods (McGrath, 2008).

The Comprehensive System (CS), developed by Exner (1974) after compiling the major elements of previous systems, provided a unified approach to using the Rorschach test over the past few decades (Exner, 2003). Although the CS is no longer evolving, in response to research demonstrating limitations in Rorschach reliability and validity, Meyer, Viglione, Mihura, Erard, and Erdberg (2011) developed the Rorschach Performance Assessment System (R-PAS). Among other revisions, the R-PAS uses a new administration procedure to reduce variation in the number of responses, focuses on variables with the strongest empirical base (Mihura et al., 2012), emphasizes the logical connection between coded behavior and inferred personality characteristic, and relies on an internationally collected normative reference group that provides percentile-based standard score transformations and also allows clinicians to adjust scores for the overall level of complexity in a protocol.

TAT. The TAT (Murray, 1943) is designed to measure drives or needs through narrative delivered by the test taker in response to a subset of the 40 available picture cards. The test taker is asked to tell a story about what is happening in the picture, with a beginning, middle, and end, including what the people pictured are thinking and feeling. Although various approaches to TAT administration and interpretation exist, typical administration entails selecting eight to 10 cards. The examiner can either use standard card sets (e.g., Groth-Marnat, 2009) or personally select cards by considering the assessment questions to be answered and the "pull" each card offers (i.e., based on typical story themes elicited by the card). Knowledge of card pull is necessary for appropriate card selection but also for interpretation of TAT stories so as to not over- or underinterpret themes that appear in the protocol. Although some scoring systems exist for the TAT (e.g., Cramer, 1990; Smith, Atkinson, McClelland, & Veroff, 1992; Westen, 1991), they are not commonly used by clinicians.

SUMMARY AND CONCLUSIONS

Although all clinicians engage in some form of clinical assessment with clients (e.g., determining appropriate interventions throughout treatment), and many make use of occasional testing (e.g., a symptom monitoring inventory), formal psychological assessment is a distinct clinical endeavor. Assessors have the unique responsibility and privilege of helping clients and others better understand the client in a way that can produce meaningful changes in the client's life. Successful assessments are achieved when the assessor ensures that he or she understands what the client and important others (e.g., referral source) want from the assessment by building a collaborative and trusting relationship with the client, making well-informed decisions when selecting and using tests, and ensuring that communication with the client and other involved parties (e.g.,

referral source, care providers, therapist) is a top priority and is conducted with respect for the client.

The field of assessment is currently in a state of transition as it has recently become a recognized clinical proficiency by the APA, and leaders in the field are working to develop the guidelines and procedures for competency evaluation. Assessment psychologists are the select few who assume the responsibility for developing and applying their knowledge of complex issues such as test validity, judgment biases, and method-related considerations to the clinical endeavor of better understanding the complexities of individual clients. Within adult mental health settings, assessors have the opportunity to help shape the perception of assessment practice in the eyes of other health professionals as an endeavor that can benefit the client in ways that therapy or medication alone cannot, and with this role comes responsibility to the field to demonstrate accurate, useful, and person-centered assessment practices.

APPENDIX 14.1 ADHD SUMMARY SCORE SHEET FOR ADULTS (CONVERT ALL SCORES TO STANDARD SCORES AND PLOT ON THE RIGHT)

ADHD Summary Score Sheet for Adults
(Convert all Scores to Standard Scores and Plot on the Right)

Initials:	**ID:**		**Age:**	**Gender:**								
Tests	**Scaled scores**	***T* scores**	**Standard scores**	Standard score profile: **Higher scores are healthy**								
Cognitive tests				**60**	**70**	**80**	**90**	**100**	**110**	**120**	**130**	**140**
WAIS-IV												
VCI												
PRI												
WMI												
Digit Span												
Arithmetic												
Letter-Number Sequencing												
PSI												
Symbol Search												
Coding												
Cancellation												
WMS-IV: Spatial Addition												
WMS-IV: Symbol Span												
D-KEFS: Trail Making Test												
Visual Scanning												
Number Sequencing												
Letter Sequencing												
Number-Letter Switching												
Motor Speed												
D-KEFS: Verbal Fluency Test												
Letter Fluency												
Category Fluency												
Category Switching: Correct												
Category Switching: Accuracy												
D-KEFS: Color–Word Interference												
Color Naming												
Word Reading												
Inhibition												
Inhibition/Switching												
D-KEFS: Twenty Questions Test												
Initial Abstraction												
Total Weighted Achievement												
D-KEFS: Tower Test												
Total Achievement												
PASAT[a]												
TOMM Trial 1/Trial 2/ Retention	**T1 = ___**	**T2 = ___**	**Ret = ____**	**Lower scores are healthy**								
CPT				**140**	**130**	**120**	**110**	**100**	**90**	**80**	**70**	**60**
Errors of omission												
RT Standard Error												
d′												
Errors of commission												
CAARS[b]												
DSM-IV Inattention												
DSM-IV Hyperactivity												
DSM-IV ADHD Total												
ADHD Index												
Informant-Report: **CAARS #1**[b]												
DSM-IV Inattention												
DSM-IV Hyperactivity												
DSM-IV ADHD Total												
ADHD Index												
Informant-Report: **CAARS #2**[b]												
DSM-IV Inattention												
DSM-IV Hyperactivity												
DSM-IV ADHD Total												
ADHD Index												

Note. WAIS–IV = Wechsler Adult Intelligence Scale—Fourth Edition; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; WMS–IV = Wechsler Memory Scale—Fourth Edition; D-KEFS = Delis-Kaplan Executive Function System; PASAT = Paced Auditory Serial Addition Task; TOMM = Test of Memory Malingering; CPT-2 = Conners' Continuous Performance Test–2; CAARS = Conners Adult ADHD Rating Scales; *DSM–IV* = Diagnostic and Statistical Manual of Mental Disorders (4th ed.); ADHD = attention deficit/hyperactivity disorder. Permission to reproduce and enlarge this score sheet is granted provided that this notice is included: "Developed by Aaron D. Upton, Wei-Cheng (Wilson) Hsiao, and Gregory J. Meyer; updated April 21, 2010. Reprinted with permission."

[a]Norm-based scores for these tests correct for education as well as age, so they should be interpreted in the context of the client's level of education. A person with relatively few years of education may score high compared with others with similar education yet still be average for their age as a whole. Also, a person with many years of education may perform poorly compared with others with similar education yet still be average for their age as a whole.

[b]CAARS ratings provided for four of eight scales. An additional four scales are available on the original CAARS forms.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research of clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341–382. doi:10.1177/0011000005285875

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological Bulletin, 137,* 211–241. doi:10.1037/a0022106

Bechtoldt, H. P. (1959). Construct validity: A critique. *American Psychologist, 14*, 619–629. doi:10.1037/h0040359

Beck, A. T., & Steer, R. A. (1993). *Beck Anxiety Inventory manual*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI–2–RF (Minnesota Multiphasic Personality Inventory—2: Restructured Form): Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion, 23*, 1021–1040. doi:10.1080/02699930802607937

Blais, M. A., & Smith, S. R. (2008). Improving the integrative process: Data organizing and report writing. In R. P. Archer & S. R. Smith (Eds.), *Personality assessment* (pp. 405–440). New York, NY: Routledge.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry and Law, 21*, 35–76.

Brewer, W. E. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 21–90). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511664014.004

Brown, N. R., Rips, L. J., & Shevell, S. K. (1985). The subjective dates of natural events in very-long-term memory. *Cognitive Psychology, 17*, 139–177. doi:10.1016/0010-0285(85)90006-4

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi:10.1037/0735-7028.31.2.141

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. doi:10.1037/1040-3590.6.4.284

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319. doi:10.1037/1040-3590.7.3.309

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*, 1092–1122. doi:10.1037/a0021212

Conners, C. K., Erhardt, D., & Sparrow, E. (1999). *Conners' Adult ADHD Rating Scales (CAARS) technical manual*. North Tonawanda, NY: Multi-Health Systems.

Conners, C. K., & MHS Staff. (2004). *Conners' Continuous Performance Test (CPT-II): Version 5 for Windows technical guide and software manual.* North Tonawanda, NY: Multi-Health Systems.

Costa, P. T., & McCrae, R. R. (2010). *NEO Inventories professional manual for NEO PI–3, NEO FFI–3 and NEO PI–R*. Lutz, FL: Psychological Assessment Resources.

Cramer, P. (1990). *The development of defense mechanisms: Theory, research, and assessment*. New York, NY: Springer-Verlag.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106. doi:10.1111/j.1529-1006.2004.00018.x

Exner, J. E. (1974). *The Rorschach: A comprehensive system* (Vol. 1). New York, NY: Wiley.

Exner, J. E., Jr. (2003). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations and principles of interpretation* (4th ed.). New York, NY: Wiley.

Finn, S. E. (1996a). Assessment feedback integrating MMPI–2 and Rorschach findings. *Journal of Personality Assessment, 67*, 543–557. doi:10.1207/s15327752jpa6703_10

Finn, S. E. (1996b). *Manual for using the MMPI–2 as a therapeutic intervention*. Minneapolis: University of Minnesota Press.

Finn, S. E. (2007). *In our clients' shoes: Theory and techniques of therapeutic assessment*. Mahwah, NJ: Erlbaum.

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). *Structured clinical interview for DSM–IV Axis II personality disorders (SCID–II)*. Washington, DC: American Psychiatric Press.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1996). *Structured clinical interview for DSM–IV Axis I disorders, clinician version (SCID–CV)*. Washington, DC: American Psychiatric Press.

Fischer, C. T. (1994). *Individualizing psychological assessment: A collaborative and therapeutic approach*. Mahwah, NJ: Erlbaum.

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin, 113,* 44–66. doi:10.1037/0033-2909.113.1.44

Ganellen, R. J. (1996). *Integrating the Rorschach and the MMPI–2 in personality assessment*. Mahwah, NJ: Erlbaum.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association. doi:10.1037/10299-000

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 254–267. doi:10.1037/0033-295X.98.2.254

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York, NY: Oxford University Press.

Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, G., Kaemmer, B., & Butcher, J. N. (2001). *Minnesota Multiphasic Personality Inventory—2: Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Groth-Marnat, G. (2009). The assessment interview. In *Handbook of psychological assessment* (5th ed., pp. 65–94). Hoboken, NJ: Wiley.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30. doi:10.1037/1040-3590.12.1.19

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY: Oxford University Press.

Hathaway, S. R., & McKinley, J. C. (1942). *The Minnesota Multiphasic Personality Schedule*. Minneapolis: University of Minnesota Press.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78–79. doi:10.1037/0003-066X.58.1.78

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureño, G., & Villaseñor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56*, 885–892. doi:10.1037/0022-006X.56.6.885

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455. doi:10.1037/1040-3590.15.4.446

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist, 43*, 614–620. doi:10.1037/0003-066X.43.8.614

Kaslow, N. J., Borden, K. A., Collins, F. L., Jr., Forrest, L., Illfelder-Kaye, J., Nelson, P. D., . . . Willmuth, M. E. (2004). Competencies conference: Future directions in education and credentialing in professional psychology. *Journal of Clinical Psychology, 60*, 699–712. doi:10.1002/jclp.20016

Koehler, J. J. (1995). The psychology of judgment and decision making: Two views. *PsycCRITIQUES, 40*, 315–316. doi:10.1037/003549

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337. doi:10.1111/j.1467-6494.1996.tb00513.x

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., . . . Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology, 60,* 725–739. doi:10.1002/jclp.20010

Kvaal, S., Choca, J., & Groth-Marnat, G. (2003). The integrated psychological report. In L. E. Beutler & G. Groth-Marnat (Eds.), *Integrative assessment of adult personality* (2nd ed., pp. 398–433). New York, NY: Guilford Press.

Lees-Haley, P. R. (1992). Psychodiagnostic test usage by forensic psychologists. *American Journal of Forensic Psychology, 10*, 25–30.

Levin, H. S., Mattis, S., Ruff, R. M., Eisenberg, H. M., Marshal, L. F., Tabaddor, K., . . . Frankowski, R. F. (1987). Neurobehavioral outcome following minor head injury: A three-center study. *Journal of Neurosurgery, 66*, 234–243. doi:10.3171/jns.1987.66.2.0234

Lichtenberger, E. O., Mather, N., Kaufman, N. L., & Kaufman, A. S. (2004). *Essentials of assessment report writing.* Hoboken, NJ: Wiley.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66. doi:10.1111/1529-1006.002

Lobbestael, J., Leurgans, M., & Arntz, A. (2011). Inter-rater reliability of the Structured Clinical Interview for DSM–IV Axis I Disorders (SCID–I) and Axis II Disorders (SCID–II). *Clinical Psychology and Psychotherapy, 18*, 75–79. doi:10.1002/cpp.693

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*, 28–50. doi:10.1177/1088868310366253

McGrath, R. E. (2008). The Rorschach in the context of performance-based personality assessment. *Journal of Personality Assessment, 90*, 465–475. doi:10.1080/00223890802248760

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minneapolis Press. doi:10.1037/11281-000

Meyer, G. J. (2010). *WIAT-III factor structure.* Unpublished raw data, Department of Psychology, University of Toledo, Toledo, OH.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Meyer, G. J., & Kurtz, J. E. (2006). Advancing personality assessment terminology: Time to retire "objective" and "projective" as personality test descriptors. *Journal of Personality Assessment, 87*, 223–225. doi:10.1207/s15327752jpa8703_01

Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. E., & Erdberg, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation, and technical manual.* Toledo, OH: Rorschach Performance Assessment System.

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2012). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the Comprehensive System. *Psychological Bulletin*. Advance online publication. doi:10.1037/a0029406

Miller, C. J., Newcorn, J. H., & Halperin, J. M. (2010). Fading memories: Retrospective recall inaccuracies in ADHD. *Journal of Attention Disorders, 14*, 7–14. doi:10.1177/1087054709347189

Millon, T., Davis, R., & Millon, C. (1997). *Millon Clinical Multiaxial Inventory—III manual.* Minneapolis, MN: National Computer Systems.

Morey, L. C. (2007). *The Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259. doi:10.1037/0033-295X.84.3.231

Norcross, J. C., Karpiak, C. P., & Santoro, S. O. (2005). Clinical psychologists across the years: The division of clinical psychology from 1960 to 2003. *Journal of Clinical Psychology, 61*, 1467–1483. doi:10.1002/jclp.20135

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33–65. doi:10.1016/j.acn.2004.02.005

Riggio, H. R., & Riggio, R. E. (2002). Emotional expressiveness, extraversion, and neuroticism: A meta-analysis. *Journal of Nonverbal Behavior, 26*, 195–218. doi:10.1023/A:1022117500440

Rorschach, H. (1942). *Psychodiagnostics: A diagnostic test based on perception.* Bern, Switzerland: Hans Huber. (Original work published in German in 1921)

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223. doi:10.1037/1082-989X.1.2.199

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206–224. doi:10.1037/1082-989X.8.2.206

Smith, C. P., Atkinson, J. W., McClelland, D. C., & Veroff, J. (1992). *Motivation and personality: Handbook of thematic content analysis.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511527937

Soldz, S., Budman, S., Demby, A., & Merry, J. (1995). A short form of the Inventory of Interpersonal Problems Circumplex Scales. *Assessment, 2*, 53–63. doi:10.1177/1073191195002001006

Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice and research. *The Counseling Psychologist, 23*, 506–534. doi:10.1177/0011000095233009

Stone, A. A., Schwartz, J. E., Neale, J. M., Shiffman, S., Marco, C. A., Hickcox, M., . . . Cruise, L. J. (1998). A comparison of coping assessed by ecological momentary assessment and retrospective recall.

Journal of Personality and Social Psychology, 74, 1670–1680. doi:10.1037/0022-3514.74.6.1670

Streiner, D. L. (2003a). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217–222. doi:10.1207/S15327752JPA8003_01

Streiner, D. L. (2003b). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103. doi:10.1207/S15327752JPA8001_18

Tombaugh, T. N. (1996). *The Test of Memory Malingering (TOMM)*. Toronto, Ontario, Canada: Multi-Health Systems.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. doi:10.1126/science.185.4157.1124

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition: Administration and scoring manual.* San Antonio, TX: Pearson.

Wechsler, D. (2009a). *Wechsler Individual Achievement Test—Third Edition: Examiner's manual*. San Antonio, TX: Pearson.

Wechsler, D. (2009b). *Wechsler Memory Scale—Fourth Edition: Administration and scoring manual*. San Antonio, TX: Pearson.

Westen, D. (1991). Clinical assessment of object relations using the TAT. *Journal of Personality Assessment, 56*, 56–74. doi:10.1207/s15327752jpa5601_6

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology, 55*, 493–518. doi:10.1146/annurev.psych.55.090902.141954

Wingenfeld, S. A., Holdwick, D. J., Jr., Davis, J. L., & Hunter, B. B. (1999). Normative data on computerized paced auditory serial addition task performance. *Clinical Neuropsychologist, 13*, 268–273. doi:10.1076/clin.13.3.268.1736

CHAPTER 15

PSYCHOLOGICAL ASSESSMENT IN CHILD MENTAL HEALTH SETTINGS

Christopher T. Barry, Paul J. Frick, and Randy W. Kamphaus

A commonly held perception is that psychological services are synonymous with treatment. Indeed, the potential of treatment to alleviate emotional and behavioral symptoms and promote adaptive functioning provides the best example of the application of psychological principles and empirical findings for the betterment of society at large. However, a scientifically grounded and comprehensive assessment is necessary to guide the design and implementation of a prevention or intervention plan. Relative to the literature on evidence-based treatment, research on evidence-based assessment is new but no less important to sound clinical practice.

This chapter presents a number of important issues involved in providing an evidence-based assessment of child/adolescent emotional, behavioral, and social functioning (i.e., psychological assessment). The chapter begins with an overview of the current state of the literature and challenges in developing an evidence base for assessing child mental health issues and describes several ethical and professional issues important in the assessment of children and adolescents. Second, the chapter provides an overview of the various tools that professionals might use in their batteries for assessing the mental health needs of children and adolescents. This section includes a discussion of the relative advantages and disadvantages of each type of assessment tool (e.g., interviews, behavioral observations, rating scales) as well as initial considerations for professionals to make when designing a battery. Third, the authors then address the unique contributions of different methods of assessment in a comprehensive assessment and provide a framework for clinicians to integrate information from multiple sources and methods. The chapter concludes with a discussion of future directions in further developing an evidence base for the psychological assessment of children and adolescents.

OVERVIEW OF EVIDENCE-BASED ASSESSMENT WITH CHILDREN AND ADOLESCENTS

A successful assessment is one that accomplishes its main goal: to answer the referral question. This kind of assessment typically involves a clear description of the types of problems a child or adolescent is experiencing and their potential causes (i.e., a case conceptualization). It also typically leads to recommendations for intervention based on this case conceptualization. Thus, the goal of an assessment is not necessarily to provide a diagnosis or to reduce descriptions of a child's functioning to a test score or set of test scores. Effective clinicians must have a good understanding of assessment techniques, psychometrics, developmental psychopathology, and state-of-the-art interventions for a host of child and adolescent behavioral, academic, emotional, and social difficulties. From this knowledge should come an assessment report that accurately and comprehensively describes a child's strengths and difficulties, provides a road map for efforts to reduce these difficulties, and is readily understandable to a variety of audiences. To aid this process, there has been a recent movement toward generating models

DOI: 10.1037/14048-015
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

for evidence-based assessment. Such efforts mirror the work over the past couple of decades on evidence-based treatment. As is the case for defining evidence-based treatment, models of evidence-based assessment must consider the diversity of settings and purposes for which psychological assessment is conducted. However, regardless of the practice setting, evidence-based assessment can provide a clear framework to guide how professionals conduct psychological assessments, communicate their findings to others, and evaluate assessment results from other professionals.

In defining evidence-based assessment, a distinction must be made between methods and processes that are evidence-based (Mash & Hunsley, 2005). Much of the focus of past discussions has been on how to use research to support the use of a particular assessment method (Garb, Wood, Lilienfeld, & Nezworski, 2005) or test (Matthey & Petrovski, 2002). Although such discussions are very important, it is a far more challenging task to determine how to define evidence supporting an approach to psychological assessment that guides the entire assessment process. The focus in this chapter includes both the methods and processes of psychological assessment of children and adolescents. The model of evidence-based assessment used herein is guided by three overarching principles:

- Every decision made during an assessment with a child or adolescent should be guided by the most current and best available research.
- Results from tests should be used only for making interpretations for which they have been validated.
- The assessment process should be guided by a hypothesis-testing approach. That is, one should address the referral question (e.g., Why is this child doing poorly in school?) by developing possible hypotheses based on research (e.g., the child has a learning disability; the child has problems sustaining attention) and then collect data to determine which hypothesis is most consistent with the available data.

The remainder of this chapter describes the practical implications of these overarching principles, starting with a discussion of how assessments guided by these principles differ from those guided by other approaches to psychological assessment.

EVIDENCE-BASED APPROACHES COMPARED WITH TRADITIONAL APPROACHES TO PSYCHOLOGICAL ASSESSMENT

One important implication of an evidence-based approach to assessment is the need to include an assessment of the child or adolescent's psychological context. That is, research consistently demonstrates the important influence of context on child development, both normal and pathological. To understand a child, it is imperative to take what Kazak et al. (2010) have referred to as a "meta-systems approach," wherein an understanding of the various systems involved with the child or available to children and families are considered in a case conceptualization and ultimate intervention plans.

In addition to being important for understanding a child's adjustment, the child's context is also important for understanding the assessment information obtained on the child's emotional and behavioral functioning. Specifically, there is a substantial body of literature indicating that ratings of a child's personality and behavior in different contexts (e.g., school and home) are only modestly correlated (Achenbach, McConaughy, & Howell, 1987; De Los Reyes & Kazdin, 2005). For example, in their meta-analysis of over 119 studies, Achenbach et al. (1987) reported that the average correlation in ratings of children's adjustment between informants who see the child in different settings (e.g., parents and teachers) was $r = .28$. In short, these modest correlations suggest that when collecting assessment information from multiple sources, it is likely that the different sources will give different views of the child's personality and adjustment. A later section of this chapter provides recommendations for integrating information across sources. However, these recommendations include trying to explain discrepancies across sources by understanding characteristics of the various contexts that may have either influenced the child's behavior in that context (e.g., an unstructured classroom without clear rules or methods of enforcement) or influenced an

informant's ratings of the child's behavior (e.g., parental adjustment problems).

Thus, it is critical that an evidence-based assessment of children and adolescents includes an assessment of the child's psychosocial context. This is in contrast to many traditional approaches to psychological assessment that focus only on describing the child's emotional and behavioral functioning—often in norm-referenced ways—but provide very little information about the family, peer, neighborhood, or cultural factors that might play a role in the maintenance, exacerbation, or amelioration of the child's problems or which may be critical for interpreting variations in the child's behavior across contexts.

Another implication of this evidence-based approach to assessment is that testing should be "construct-centered," as opposed to diagnostic-centered or test-centered. As noted earlier, the first principle of the model of evidence-based assessment used herein suggests that knowledge of current scientific findings regarding specific assessment issues, as well as about child development and psychopathology, should inform the assessment process. For example, research clearly suggests that comorbidity, or the co-occurrence of different types of problems, is the rule, not the exception (Bird, Gould, & Staghezza, 1993). Therefore, to focus solely on a specific diagnosis would miss a host of factors that may also be present or that most certainly influence the child's presentation and functioning. Diagnostic systems are important insofar as they facilitate communication between professionals and also help convey (i.e., to a school system or to a third-party payer) the appropriate level of services for a child. However, diagnostic systems also are imperfect and their misuse can have deleterious effects on a child.

Test-centered approaches are also problematic. Tests that yield scores are imperfect and are designed to describe only a certain aspect of the child's functioning. Considering a child as fully described by his or her performance on a single test misses much critical information. For example, elevated scores on multiple domains of a broadband rating scale does not necessarily mean that multiple diagnoses or targets of treatment are appropriate for a child. Likewise, a lack of elevations on a rating scale does not mean that there are not issues for which intervention could be beneficial. In short, diagnoses and test scores should not be the centerpieces of assessment results. Instead, focusing on descriptions of primary and secondary difficulties (which may indeed warrant a diagnosis and may be evidenced by test scores), their apparent underpinnings, and recommended interventions will result in an assessment that is potentially of great benefit to the child and his or her family. Research in developmental psychopathology can guide this process. For example, the literature on manifestations of child/adolescent depression directly informs the important constructs to assess for children referred for depression, which would not only include appearing sad but also could include somatic complaints, loss of interest in activities, thoughts of death, and psychomotor retardation during adolescence (Weiss & Garber, 2003).

As noted earlier, another principle of an evidence-based approach to assessment is that the assessment process is one in which the clinician engages in hypothesis testing and arrives at conclusions that inform interventions based on the data collected during the evaluation (see also Ollendick & Hersen, 1993). Thus, an assessment report should be similar to a scientific manuscript. First, much like a literature review and presentation of a research question, the report includes a statement of the reason for referral and a description of the client's background information that describes the primary presenting problem and the goal of the assessment. A list or description of the assessment methods used in the evaluation is analogous to a Method section in a scientific article and provides the reader with an understanding of the tools used to draw conclusions about the referral question. Results are presented, and in true scientific fashion, preference should be given to results that are couched in measurable terms rather than being based solely on the clinician's interpretations. From there, much like a Discussion section in a manuscript, the report writer interprets the findings on the basis of the actual results of the assessment. This discussion should make conclusions succinctly as to how the data (i.e., test results) seem to address referral questions. More important, it clearly outlines which hypotheses to explain the referral question are most

consistent with the test results. Many journal articles conclude with a discussion of future directions in a particular area of research and recommendations offered in a report can be thought of in the same way. In light of the client's presenting problem and the results of the assessment, recommended strategies to address the problem are presented to interested parties (parents, teachers, etc.). In such an approach to assessment and report writing, the clinician is acting as an applied scientist by developing a theory (case conceptualization) of the client based on gathered data (see Frick, Barry, & Kamphaus, 2010). Readers also may consult Chapter 3 in this volume.

ASSESSMENT OF TREATMENT OUTCOME

Much of this chapter is focused on assessments primarily designed to guide the development of a treatment plan. It should be noted, however, that psychological assessment can involve continued progress monitoring during the course of treatment. Little information exists as to the frequency with which clinicians formally evaluate the effectiveness of their treatments either during the course of treatment or at the end. What is known is that regular assessment of change during treatment increases treatment fidelity and improves treatment outcomes (Lambert et al., 2003); thus, this form of assessment in child mental health settings should become routine. Readers interested in assessment in adult mental health settings are referred to Chapter 14 in this volume. Readers seeking additional information about assessment of treatment outcomes in medical settings are directed to Chapter 18 in this volume.

An overarching model of evidence-based assessment can be used to guide this type of assessment as well. First, the criteria by which treatment progress is evaluated should be measureable, which is not to say that the data must necessarily be numerical. For example, ordinal ratings (e.g., "never," "sometimes," "often") of the frequency of a target behavior can provide useful information about relative change in the behavior. Second, only measures that have proven to be sensitive to change should be used for the purpose of treatment monitoring. For example, the response scale on a parent-report behavior rating scale may be too general (e.g., "never" vs. "sometimes" vs. "always") or the time interval for reporting the frequency of a parent behavior (e.g., the past 6 months) may not be discrete enough to detect changes brought about by treatment (McMahon & Frick, 2005). Third, the criteria for evaluating treatment outcome should be meaningful, which can be defined in terms of normative functioning but most often should be defined in terms of the child's relative functioning. To establish meaningful outcomes, baseline data on the referral issues of concern are essential. Fourth, the criteria for evaluating treatment outcome must be feasible, which again is client specific. For example, it may very well be meaningful to consider progress to be reflected by a reduction of incidents of talking back to parents from 30 such incidents daily to zero each day by Session 6, but such a dramatic elimination of this type of behavior may be unrealistic. Similarly, to propose that treatment progress be defined as a reduction of problems from a clinically significant range to a normal range on a rating scale would indeed indicate clinically significant progress. However, for many children for whom clinical services are needed, a normal level of functioning after treatment may not be a reasonable goal. Instead, incremental, meaningful improvements may not only be more feasible, but more important, such treatment gains also may reduce the child's impairment.

ETHICAL AND PROFESSIONAL ISSUES IN THE ASSESSMENT OF CHILDREN AND ADOLESCENTS

In conducting research, the scientific demands of a study must always be secondary to a number of important ethical and professional issues. The same is true in psychological assessments, and there are several unique ethical considerations in the assessment of children and adolescents. Successful execution of the assessment can be facilitated and ethical issues avoided through appropriate planning. As part of this process, the clinician should first determine whether an evaluation is warranted and whether he or she is suited to conduct it (Frick et al., 2010). A critical part of planning is to determine who has the right to consent for the assessment to

be conducted and to communicate to relevant parties the potential need to have documentation to that effect (e.g., for divorced parents, the mother brings paperwork to the assessment documenting that she has legal custody and, therefore, the right to seek services).

Additionally, the clinician must be effective in establishing and maintaining rapport with parents, children, and teachers. Everyone who will provide information in the assessment should be informed about what to expect and how information they provide will be used. This information, of course, would be relayed to a child in developmentally appropriate language. Moreover, because the child or adolescent is typically not self-referred, his or her understanding of the evaluation or motivation may be limited, making efforts to explain the assessment process in clear, reassuring ways all the more necessary.

Similarly, expectations for confidentiality should be discussed with all participating parties at the outset. It is to be expected that relevant information from all informants will be disclosed in an assessment report, yet the clinician should take care to provide only information relevant to the purpose of the assessment in the report. It is important that parents and children understand to whom information may be released at the end of the assessment with the parent having an opportunity to consent to that release after the report is completed. There are, of course, instances in which confidentiality cannot be ethically or legally maintained. Ethical and legal obligations to report abuse or neglect and to report concerns regarding harm to self or others should be discussed before the assessment starts. All parties, including children, should be able to understand the clinician's explanation of the situations in which disclosing information might be necessary and should be allowed to ask questions regarding this issue. Regardless of the types of tests to be administered or the apparent focus of the assessment, all professionals should discuss the limits of confidentiality at the outset of any assessment and should be prepared to reiterate these limits if a client, parent, or teacher seeks reassurance during the assessment that information provided will be confidential.

Although many scenarios could be presented, the key point is that it is incumbent upon professionals who conduct assessments to do so in an ethical and competent manner. Elsewhere, Frick et al. (2010) provided a self-examination that a professional can use in determining if his or her involvement in, and conduct of, the assessment meets this obligation. In short, professionals should

(a) ensure that they have appropriate training for the assessment methods to be used,
(b) consider the client's background in interpreting assessment results,
(c) receive informed consent before initiating the assessment,
(d) consider to whom assessment feedback should be provided,
(e) take appropriate steps to maintain the client's confidentiality, and
(f) obtain releases to provide information from the assessment to outside parties.

A professional who routinely addresses each of the aforementioned issues will likely avoid many potential professional or ethical pitfalls. It is important to note that the preceding recommendations should be considered the minimum necessary to conduct an effective assessment in an ethical and professional manner. Also, much of this discussion of an evidence-based approach to assessment has focused on the *process* of assessment. This focus is important because it is often neglected in discussions of psychological assessments. However, it is still important to consider the implications of an evidence-based approach to assessment for the *methods* used in conducting assessments of children and adolescents.

ASSESSMENT METHODS AND MEASURES FOR USE WITH CHILDREN AND ADOLESCENTS

General Issues in Selecting Measures

Psychological assessments of children and adolescents necessitate the use of multiple methods of gathering information on the construct(s) of interest (Kazdin, 2005). Fortunately, a great deal of research has been conducted on the reliability and validity of specific measures of a variety of clinical constructs. On the one hand, clinicians can use this research to

guide decisions as to which measures will be most appropriate for a given case. On the other hand, however, very little evidence exists that provides a clear framework for how specific measures together might yield a comprehensive, yet parsimonious, approach to answer a referral question or how various assessment methods should be differentially weighted in clinical decision making. In short, an empirical foundation exists for the initial selection of measures, but the available evidence is much more limited in how a clinician should appropriately integrate information provided by the chosen tools.

Selection of methods and measures must take into account that the meaning of a child's presenting difficulties is based partly on the child's developmental context. For example, the same behavior (e.g., substance use, hyperactivity, defiance or oppositional behavior) takes on a different meaning if the child is 2, 10, or 16 years old (Frick et al., 2010). Whether a behavioral problem is atypical for a developmental context or represents an exaggeration of a more typical developmental process is a critical factor for case conceptualization. Likewise, assessment measures should allow the clinician to discern between developmentally normal and abnormal functioning for the child. In addition, for normative comparisons to be made, the measure must have adequate norms to compare a child client's presentation with that of his or her same-aged peers (Frick et al., 2010).

Of course, basic psychometric characteristics (e.g., reliability, validity, norm sample) of a measure are important in deciding whether it should be part of an assessment battery. That is not to say that the battery must consist exclusively of a series of tests that meet a threshold of acceptable psychometric properties. Indeed, client-specific information not captured by a measure as well as clinical skill and clinical judgment, are important aspects of the assessment process, as the clinician is charged with integrating multiple sources of information from multiple informants into a coherent case conceptualization that lends itself to meaningful and feasible intervention strategies for a particular child. Professionals, nevertheless, need to consider the appropriateness of any assessment measure based on the purpose of the evaluation and the child's developmental level.

Content, reliability, and validity are certainly part of this process. A critical consideration in this process is the fact that tests themselves are not "reliable" or "valid." Instead specific uses of test scores can yield reliable or valid results that other uses may not. For example, many structured diagnostic interviews have demonstrated acceptable levels of reliability when used with children over the age of 9 but are less reliable when used with younger samples (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985). Similarly, certain interpretations of test scores may be valid, whereas other interpretations may not. For example, scores on the Aggressive Behavior scale of the commonly used Child Behavior Checklist have been correlated with other measures of conduct problems; the correlations that result offer evidence supporting the use of these scores as a screener for a diagnosis of either oppositional defiant disorder or conduct disorder (Achenbach & Rescorla, 2001). However, the content of the scale includes a large number of items that are not specific to physical aggression (e.g., argues a lot, mood changes). Therefore, it would be appropriate to use scores on this subscale as a norm-referenced indicator of the level of a child's conduct problems but not to use the scores to determine whether a child shows higher rates of physical aggression than other children. When selecting measures for an evaluation, an assessor must consider whether the scores from the test have proven to have acceptable reliability in the population for which he or she wants to use it and whether there is evidence to support the validity of the interpretations he or she would like to make from the test scores.

Another important issue in selecting measures is evaluating the clinical utility of a particular tool. According to Mash and Hunsley (2005), clinical utility involves the extent to which a measure "will make a meaningful difference in relation to diagnostic accuracy, case formulation considerations, and treatment outcomes" (p. 365). A related notion is incremental validity, or the extent to which the addition of a measure provides unique additional information that aids in the assessment process (Johnston & Murray, 2003). These concepts can be applied to a battery or set of measures in that an additional set of measures or an additional

assessment may or may not improve understanding of the child's difficulties. If incremental validity is absent, the need for further assessment using the measures in question is also absent. Unfortunately, limited research has been conducted on the clinical utility or incremental validity of many measures.

Finally, assessment techniques vary in their utility for certain interpretations. Stated another way, there is no such thing as the perfect test. All techniques have certain strengths and weaknesses in what they add to an assessment battery. Thus, this consideration supports the need for an assessment battery, rather than use of a single test, when providing a psychological assessment of children and adolescents. It also suggests that, when designing an assessment battery for a child or adolescent, it is critical that the design consider the unique contributions of different methods of assessment. In the following sections, some of the most common assessment methods are evaluated and some of the key considerations are highlighted that should guide clinicians in determining if and how the different methods should be used in an assessment battery.

Clinical Interviews

Historically, a critical part of an assessment battery is the clinical interview with the child, his or her parent, and with other important adults who interact with the child (e.g., teacher). Most commonly, this interview involves taking a history of the presenting problem, gaining a description of the problem and the impairments that it appears to cause, and gaining information across domains of functioning on factors that might influence the manifestation of the problem. Because such interviews are, by nature, unstructured and idiosyncratic to the client (and interviewer), they are often unreliable. In other words, the interview will not be conducted the same way across clients or across assessors. Still, unstructured clinical interviews provide invaluable information about the client's particular history, problems, and strengths, but they do not allow for conclusions about the extent to which the child's difficulties are significant relative to same-aged peers. Instead, they set the stage for further assessment activities in that allowing the caretaker to articulate his or her concerns helps the clinician determine specific issues in need of further evaluation. For example, an unstructured interview may inform the clinician that symptoms of attention-deficit/hyperactivity disorder (ADHD), the history of these symptoms, and their pervasiveness should be emphasized as the evaluation continues, whereas, on the basis of the interview, mood symptoms should not.

The flexibility and client-centered nature of unstructured clinical interviews make them ideal for determining important features of the child's presentation such as the onset of the problem, the relation of the problem to significant environmental events or stressors, the course of the problem, the child's previous assessment or treatment history, and family psychiatric history (Frick et al., 2010). That is not to say that the clinician's approach to an unstructured interview should be haphazard or that such an interview should not be guided by research. Instead, like all parts of the assessment process, the interview should be guided by the most recent research on development and psychopathology. For example, if a clinician discovers that, in childhood, the client showed significant hyperactive behaviors but that these behaviors have seemed to lessen in severity in adolescence, whereas the inattentive behaviors remain problematic, it is important for the clinician to know that research suggests that this is a very common developmental history for children with ADHD (Lahey & Willcutt, 2010).

Structured diagnostic interviews are interviews that provide a specific script for the interviewer to follow while still covering relevant symptomatology, onset, and impairment related to the symptoms. That is, these types of interviews provide an opportunity to evaluate many important elements of the child's symptoms. They also include explicit guidelines on how a child's responses are to be scored. Such interviews are generally structured around stem questions (e.g., "Have you been involved in many physical fights?"), followed by a series of follow-up or contingency questions to define relevant parameters such as frequency (e.g., "How many fights have you been in the past year?"), severity (e.g., "Have you ever used a weapon in a fight?"), duration (e.g., "When was the first time you got in trouble for fighting?"), and impairment (e.g., "Has fighting caused problems for you at school, home, or

with kids your age?"). Because of the stem and follow-up format, the length of time that it takes to administer a diagnostic interview is heavily dependent on the number of problems being experienced by the child, with most interviews taking between 60 and 90 minutes to administer. Examples of widely used structured interviews that assess a variety of diagnostic categories include the Diagnostic Interview Schedule for Children–4 (DISC-4; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) and the Diagnostic Interview of Children and Adolescents (DICA; Reich, 2000), both of which include formats for parent and child reports.

The structured and standard formats lead to information that has shown to be more reliable than what is obtained from unstructured interviews. In addition to reliability, many structured interviews clearly have content validity, as their content is often derived directly from diagnostic criteria (Frick et al., 2010). Thus, structured diagnostic interviews can be very helpful in determining whether the child meets criteria for a particular diagnosis. On the other hand, structured interviews have a number of disadvantages, including the amount of time required to administer the interview, the reliance on the informant's subjective report as to whether a symptom is present and when it first emerged, the lack of direct translation between meeting diagnostic criteria, and the most appropriate interventions for a specific child. Most important, norm-referenced interpretations (i.e., determining whether the level of anxiety is more than what is normative for children of the same age) are typically not possible from diagnostic interviews (Frick et al., 2010). It is also important to recognize that many issues that might be worthy of clinical attention in children do not fall neatly within a diagnostic category (e.g., problems in peer relations), making sole reliance on diagnostic interviews unwise.

Behavioral Observations

Another important part of the assessment of children and adolescents are behavioral observations of the child either during testing, in interactions with their parents, in the classroom, or all of the above. In fact, some have historically viewed behavioral observations as the criterion against which the validity of other assessment tools should be measured (see Frick et al., 2010). Behavioral observations, by nature, provide information on a child's behavior that is not filtered through the perspective of an informant. Also, behavioral observations can be conducted in a naturalistic setting (e.g., classroom), allowing for an understanding of the environmental factors that might influence or even trigger the child's behavior. Without behavioral observations, a clinician might have some information on the presence of certain child behaviors (e.g., aggression) but might not be aware that there are specific antecedents (e.g., teasing by peers) that tend to elicit the behaviors of interest (see Carroll, Houghton, Taylor, West, & List-Kerz, 2006). The intervention implications of such environmental influences could be quite important (e.g., teaching the child how to respond to teasing in a nonaggressive manner).

Behavioral observations can be conducted informally based on the observer's descriptions of the child's behavior or through available structured observational systems, such as the Behavior Assessment System for Children (2nd ed.)–Student Observation System (BASC-2-SOS; Reynolds & Kamphaus, 2004) and the Test Observation Form (TOF; McConaughy & Achenbach, 2004). If a clinician opts for an informal approach, it is still important to capitalize on the relative strengths of observations by noting the features of the environment in which the observation takes place, the specific behaviors of note, the antecedents to those behaviors, and the responses of others in the environment (i.e., the consequences) to the behavior. Structured systems often prescribe parameters such as the approach to coding (e.g., time sampling, event recording, duration), the target behaviors to be coded, and the number of observation periods. Often, these systems call for the use of more than one observer as well as the observation of another child to offer a comparison between behavior exhibited by the target child and another child's behavior in the same context (McConaughy & Achenbach, 2004). Consequently, there may be a trade-off between the thoroughness or structure of the observation and its efficiency or cost effectiveness (Frick et al., 2010).

The clinician must also consider reactivity on the part of the child being observed as well as others in

the observation setting. To address this issue, the clinician should consult with others in the observation setting (e.g., teachers) about the best time to conduct the observation and how to best avoid disrupting the setting. The clinician should also consider who best to conduct the observation, particularly if the child has had contact with the clinician previously. Last, older children may be more likely to recognize the presence of an observer and to alter their behavior accordingly, making behavioral observations potentially more useful for younger children (Frick et al., 2010). Even without behavioral observations in outside settings, observations during testing regardless of the client's age and should be incorporated into the assessment report.

The use of naturalistic behavioral observations in child assessments is essentially an issue of trade-offs. Behavioral observations in a naturalistic setting provide a level of ecological validity offered by no other method, yet the clinician does not have control over the environment in the way that he or she has control over the scope or focus of an interview or an observation in a clinic setting. Behavioral observations provide direct collection of relevant evidence on the child's behavior but miss many relevant internal states (i.e., cognitions, emotions) that play a role in the child's functioning. The clinician is able to get a depth of information on the interplay between the child's behavior and his or her context, but doing so takes time and resources. Like clinical interviews, behavioral observations of some type are necessary for assessments of children, but they typically require complementary methods to overcome their limitations.

Tests of Intellectual Functioning and Academic Achievement

If a main task of child assessments is to understand the child's difficulties and strengths within the context of his or her developmental level, then a key piece of information in many cases is the child's current cognitive or intellectual functioning. Because intelligence tests have long been used in psychological assessments of children and adolescents, there are decades of research devoted to the constructs that make up intelligence, the correlates of intellectual test scores, and the interplay between and among intellectual, behavioral, emotional, social, and academic functioning (see Kamphaus, Reynolds, & Vogel, 2009). Well-normed standardized intelligence tests such as the Wechsler Intelligence Scale for Children (WISC; Wechsler, 2003) and Stanford–Binet (Roid, 2003), to name just a couple, have the advantage of clear procedures for administration and scoring. They provide unique and important information in terms of the level at which the child's verbal and nonverbal reasoning abilities have developed. These tests carry the disadvantages of requiring specialized training to administer and score and taking more time than many other assessment techniques. However, measures of intellectual functioning can be critical for understanding a child's adjustment and important for treatment planning in a number of ways. For example, research clearly indicates that intellectual functioning is a critical consideration in the design of interventions for youth with autism spectrum disorders (see Ozonoff, Goodlin-Jones, & Solomon, 2005) and that intelligence, particularly verbal intelligence, influences the manifestation of child conduct problems (Loney, Frick, Ellis, & McCoy, 1998).

Depending on the referral question, an evaluation may also include a standardized test of academic achievement. Traditionally, learning disability evaluations included these tools so that a direct comparison to intellectual functioning could be made, with a learning disability being defined as significantly lower academic achievement relative to one's measured intellectual ability (American Psychiatric Association, 2000). However, this approach to assessing learning disabilities has been questioned because of its inherent assumption that the factors that negatively affect academic achievement scores do not influence scores on intelligence tests and because classification of a child is based on static performance rather than performance on academic tasks over time. This approach is inconsistent with how many school districts determine the need for academic intervention, which often involves a multitiered evaluation of the child's response to increasing levels of intervention in the area(s) of academic difficulty (see Fletcher, Francis, Morris, & Lyon, 2005). Otherwise, standardized academic achievement tests can provide an important metric of the

impairment that is presumably caused by a child's attention, behavioral, or emotional problems. That is, if a main referral concern has to do with attention problems, the child's relatively lower achievement might be conceptualized as an effect of these attention problems in light of other converging evidence. A clinician must weigh the advantages of obtaining this information in light of the referral question and the time that it takes to administer achievement tests properly.

Behavior Rating Scales

Behavior rating scales have become a centerpiece of child psychological assessments because of their convenience, their ability to assess a large number of domains relevant to a child's psychological adjustment, and their typically sound standardization and norming processes that easily allow for age-based comparisons on constructs of interest. Concerns about potential reporting biases in rating scales are legitimate, particularly when there may exist motivation by some informants to underreport symptoms, limited opportunity for some informants to observe relevant behaviors, or a tendency for informants who have a negative view of the child's behavior to overreport problems in multiple domains (Frick et al., 2010). A recent advance in child behavior rating scales that follows from adult personality assessment is the inclusion of validity scales. These scales typically are meant to capture tendencies to present the child in an overly positive or negative light, inconsistency in responses across similar items, and a tendency to respond carelessly (Frick et al., 2010).

Broadband, or omnibus, rating scales are those that have a number of subscales assessing different domains of functioning (e.g., conduct problems, anxiety, social skills, depression, adaptive behavior). Some examples of current and widely used rating scale systems include the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001), the Behavior Assessment System for Children–Second Edition (BASC-2; Reynolds & Kamphaus, 2004), and the Conners–3 (Conners, 2008). Each of these systems provides parent-, teacher-, and self-reports. The issues to be considered for each of these informants are discussed later. Parent and teacher rating scales are often closely aligned in content and response format, and current widely used rating scales such as those listed earlier typically have extensive information supporting their reliability and criterion-related validity (i.e., the extent to which subscales correlate with other measures of the same construct).

Omnibus rating scales have historically focused on evaluating the presence of problems in adjustment. However, many of the more popular rating scale systems such as the BASC and ASEBA have recently incorporated more extensive assessment of adaptive functioning with parents and teachers as informants. In addition to being included on omnibus rating scales, adaptive functioning also can be evaluated through stand-alone rating scales or interviews with parents or teachers that are designed to provide an extensive, norm-referenced indication of adaptive functioning. One example of this type of measure is the Vineland Adaptive Behavior Scales (Sparrow, Cicchetti, & Balla, 2005). There have been clear advancements in the availability of standardized tools to measure adaptive functioning, but again, the clinician must determine the cost-effectiveness of including particular measures for this purpose. The advancements in this area are nevertheless welcome, as the assessment of adaptive functioning, including in areas such as functional communication, social skills, and self-care, has implications for the design of interventions for many young people and is a necessary component of evaluations for developmental delays and mental retardation.

Several improvements have been made over recent years in the available child self-report rating scales (see Frick et al., 2010, for a review). First, many of the commonly used child self-report inventories are closely aligned with parent and teacher counterparts, potentially facilitating interpretation of primary areas of concern. Second, many child self-report inventories (e.g., the Minnesota Multiphasic Personality Inventory—Adolescent; MMPI–A; Butcher et al., 1992) were derived from adult inventories that have seen extensive use and research over many years. The MMPI–A contains, in the MMPI tradition, clinical scales from which profiles can be derived as well as content scales that cover a variety

of domains, particularly those involving mood and affective symptoms. However, it is quite long, relative to other available rating scales. The self-report versions of the ASEBA (Achenbach & Rescorla, 2001) and BASC-2 (Reynolds & Kamphaus, 2004) sample numerous clinical and adaptive domains, whereas the Conners-3 (Conners, 2008), like its parent- and teacher-report forms, provides an extensive evaluation of symptoms of ADHD and other disruptive behaviors. Because the updated versions of the ASEBA, BASC, and Conners are relatively new, comparatively little research exists concerning their validity and clinical utility.

Overall, omnibus rating scales, particularly those that cover theoretically relevant domains and have good psychometric properties, generally have the advantage of providing norm-referenced information in a reliable and cost-effective manner (Frick et al., 2010). However, information from rating scales is filtered through the perspective of an informant, and they lack the depth of client-specific information necessary to ultimately arrive at an individualized case conceptualization. Knowledge of some of the potential pitfalls of using rating scales should greatly assist the clinician in selecting rating scales and in appropriately integrating their results with other available findings.

The emergence of a number of single-domain rating scales over the past several years is consistent with changes in assessment practice in general. Mash and Hunsley (1997) noted that assessment activities have tended to become more disorder or problem specific, have tended to emphasize the inclusion of relevant content (i.e., specific symptoms), and have been oriented toward efficiency in terms of time and focus. Therefore, single-domain rating scales appear to have an important place in current assessment practice. Single-domain rating scales typically follow the response format of broadband rating scales in that the informant is asked to endorse the presence and/or frequency of a particular problem or characteristic for the child. Single-domain rating scales may have items that directly align with diagnostic symptoms and also may include secondary factors that are often seen in individuals with a particular problem or disorder (e.g., social withdrawal for individuals with depressive symptoms).

Because single-domain scales are geared toward the evaluation of a specific problem, they are not as widely used and often not as widely known as broadband scales. It is interesting that rating scales of internalizing problems such as the Child Depression Inventory (Kovacs, 1992) and the Revised Children's Manifest Anxiety Scale (Reynolds & Richmond, 1985) have a long-standing history of use, and the evidence to support their use is much greater than what is available for assessing externalizing problems. McMahon and Frick (2005) reviewed a few such scales for the assessment of conduct problems but also noted that these scales tend to be limited only to certain aspects of conduct problem behaviors, perhaps because of the myriad ways that externalizing problems can manifest.

The obvious appeal of these single-domain, or narrow-band, rating scales is that they allow for greater depth of information in a time-efficient manner on an important construct. A problem with relying on single-domain rating scales is that they do not include items evaluating the presence of other issues. Therefore, if a clinician suspects that other issues are relevant, he or she is faced with the decision to include additional single-domain rating scales, go back and administer an omnibus rating scale, or search for other methods for assessing the additional problems. The routine use of broadband rating scales, augmented by a more in-depth assessment of the primary issue(s) of concern by means of a single-domain rating scale or interview, would circumvent this problem.

Laboratory Tasks

Laboratory tasks constitute another method for assessing psychological functioning. Such tasks are designed to elicit performance that will help confirm or disconfirm the presence of a specific problem. The tasks are based on theoretical ideas of how an individual with a particular problem (e.g., ADHD) would behave in a contrived situation. Thus, if a child truly has the underlying difficulty, his or her behavior or performance on an analogue task presumably would be consistent with how others with the same problem would respond to the task. For example, the Continuous Performance Test (CPT; Conners, 1995) is often used in the assessment of

ADHD. This task requires that respondents press a computer key if a particular letter is flashed on the computer monitor and to inhibit the response if any other letter is shown. From this task, errors of commission, which can be indicative of poor impulse control, and errors of omission, which can be indicative of inattentiveness, are recorded. An index is then derived that yields a probability that the child behaved in a manner consistent with children with ADHD. Other analogues include behavioral avoidance tasks for assessing anxiety (see March & Albano, 1996), lexical decision tasks to evaluate how quickly respondents respond to affect-laden words (Loney, Frick, Clements, Ellis, & Kerlin, 2003), and reward dominance tasks wherein the individual's response on a game to increasing penalties for previously rewarded responses is assessed (e.g., O'Brien & Frick, 1996).

Some evidence supports the potential utility of laboratory tasks for the assessment of response to treatment for individuals with conduct problems (see Frick & Loney, 2000), whereas the validity of such tasks for assisting with actual diagnostic decisions for problems such as ADHD is questionable in that performance on these tasks, for example, may not differentiate between youth with and without ADHD in the way that direct assessment of symptomatology would (see Pelham, Fabiano, & Massetti, 2005). Thus, performance-based tasks should not replace the other elements of a comprehensive assessment. Instead, like any other source of information, performance on a laboratory task should be integrated with the results of other techniques and reports from informants.

Benefits and Challenges of a Comprehensive Assessment Battery

Several findings from research on childhood emotional and behavioral problems have important implications for the assessment process. First, it is clear that a child's emotional and behavioral functioning may vary across different situations (e.g., home vs. school). Second, it is also clear that children with problems in one domain (e.g., anxiety) are likely to have problems in others areas of adjustment (e.g., depression, peer relations). Finally, there is not a single best method for assessing all of the important constructs that contribute to understanding a child or adolescent's emotional and behavioral functioning. Therefore, it is essential that an assessment battery includes procedures that provide data from multiple informants who interact with the child in different settings and who may have different perceptions of the child's adjustment. Furthermore, it is important to use different methods so that the strengths of one method can compensate for limitations in another.

Thus, psychological assessments for children and adolescents must utilize multiple assessment methods from several different informants. Unfortunately, this approach leads to one of the most challenging aspects of conducting psychological assessments of children and adolescents; namely, how to integrate many different sources of information into a clear case conceptualization that addresses the referral problems and points the way to the most effective treatment for the child or adolescent. Research on the advantages and disadvantages of assessment information from various informants and methods can guide the clinician in this difficult but important process.

Parent informants. For children before adolescence, a parent is thought to be the most useful and critical informant (Frick et al., 2010). There is some evidence that parent reports may become less useful as the child gets older (Paikoff & Brooks-Gunn, 1991), but, at the very least, the parent can still provide a developmental history. At most, the parent can continue to serve as a source on many areas of the child or adolescent's functioning in a manner that informs treatment. For example, for an adolescent with suspected behavioral problems, the parent's perception of the severity or frequency of such behaviors, and the degree of convergence with the adolescent's own self-report, may inform the clinician about the parent's monitoring of the adolescent and the possibility that increasing parental monitoring or supervision would be an appropriate target of intervention.

Research provides some guidance as to factors that might affect the validity of parental reports and that should be considered by the clinician conducting the assessment of the child. Parental depression or psychopathology, for example, may influence a

parent to view the child in a particularly negative light across domains of functioning (Richters, 1992). In addition, parent ratings may be influenced by the view that the parent takes regarding the cause of the child's difficulties (i.e., dispositional vs. situational; De Los Reyes & Kazdin, 2005). Despite these issues, parents should be viewed as vitally important for obtaining information about a child's history and description of current functioning. Even if parent reports are influenced by factors not related to the child's actual functioning, the source of this influence may be useful for case conceptualization. For example, even if a parent's level of depression leads to a negative view of a child's adjustment, such negative perceptions are likely to also influence parent–child interactions in the home, and this may be important for designing family-based interventions.

Teacher informants. Through much of childhood and adolescence, a child may spend more time in school than in any other single setting. The school setting provides many demands (e.g., to stay seated, to interact with a large number of peers, to follow rules) that may not be present to the same degree in other settings. As a result, many emotional and behavioral problems are most evident and cause the greatest level of impairment at school. Therefore, obtaining information from teachers is often vital in child and adolescent psychological assessments.

Like any informant, there are limitations in the information provided by teachers. Whereas teachers typically are good informants concerning attention problems and hyperactivity based on the unique demands of the classroom setting, they often have less of an opportunity to observe some forms of antisocial behavior (e.g., fire setting, stealing) or internalizing problems (e.g., anxiety; Loeber, Green, Lahey, & Stouthamer-Loeber, 1991). In addition to presenting problems, the age of the child also influences the usefulness of teacher reports in that an individual teacher has more of an opportunity to interact with and observe a younger child than is typically the case for adolescent students who often have a large number of teachers throughout the school day (Edelbrock et al., 1985).

It is important to note that teachers are in a unique position of interacting with many children at a particular age or developmental level; thus, they have a normative reference against which to compare the child client. This knowledge of typical child functioning or behavior could extend over many years of experience; however, the specific population with which the teacher has worked (e.g., children receiving special education services) is an important consideration for interpreting the teacher's normative perspective (Frick et al., 2010).

Child informants. Children and adolescents can provide useful information on some clinical constructs, particularly covert conduct problems (e.g., lying, stealing) and internalizing symptoms that may be unknown to other informants (Frick et al., 2010). As noted earlier, children and adolescents are becoming more central participants in assessment activities, yet their motivation to participate and provide information may be suspect because they typically are urged to get an evaluation by someone else. Traditionally, concerns have been raised about social desirability influencing children's reports; however, research indicates that the extent or direction of this influence, particularly concerning the correspondence between child and parent informants is unclear (De Los Reyes & Kazdin, 2005). If the young person is reasonably engaged in the assessment process and is able to comprehend the questions being asked of him or her, then the clinician should be able to place confidence in the validity of the self-report.

As with information from any informant, self-reported behavior and symptomatology should be integrated with that obtained from other sources. In fact, it may not be the child's actual report that is most relevant for his or her functioning; instead, the child's perception of the construct being evaluated may reveal relevant information about the case. For example, relative to parental reports of parenting practices, youth perceptions of their parents' parenting have been found to be more closely related to the youth's behavioral problems (Barry, Frick, & Grafeman, 2008).

Peer informants. Peer-referenced assessment is an intriguing, yet rarely used, method in child assessments. This approach involves ratings of the child by his or her peers through nominations by a group

of peers (e.g., classmates) on criteria of interest (e.g., fights most, most friendly, liked most). The most convenient setting in which to use peer informants is the classroom. Because of limited access, time, and ethical concerns about engaging a group of peers in an assessment of a particular child, the feasibility of peer informants is significantly constrained. However, peers provide a unique perspective on the child's social functioning and may reveal interpersonal issues that inform intervention efforts. For example, if a child has ADHD and experiences social rejection (as indicated by peer nominations), the child's peer relationships might be a target of treatment in addition to typical strategies designed to assist the child in managing his or her symptoms of ADHD.

In determining whether to use peer informants, the professional must take care to minimize the disruptiveness of the process to the peer group's normal routine (e.g., in the classroom) and should make every effort to ensure that peer informants are not aware of the target of the assessment and that they understand the importance of keeping their responses confidential even after the procedure is complete (Frick et al., 2010). Research has demonstrated that a subset of individuals from a peer group or classroom can provide nominations that are highly correlated with those obtained by the entire class (Prinstein, 2007). Thus, one way to obtain peer reports while managing the potential drawbacks of this method might be to obtain nominations or ratings from a relatively small group of peers. Unfortunately, thoughts on the usefulness of peer informants are theoretical and speculative at this point, as no empirical data exist as to the incremental validity of peer reports within assessments that include other informants and methods.

Institutional records. Another source of behavioral, academic, and social information are school or institutional (i.e., residential treatment center, detention center) records. These records may include grades in school, disciplinary infractions, incidents with peers, or awards. In the case of documented problems, records provide a clear indicator of impairment in the setting from which records are obtained. For example, it is one thing for a parent or teacher to report that a child is frequently in trouble at school, but it is another to see that the behavioral problems have risen to the level of formal documentation and disciplinary action. Likewise, concerns about the child's academic performance cannot be truly understood in the context of his or her academic demands without some indicator of performance at school, such as test scores and grades. Unfortunately, there is no clear empirical evidence as to the validity or utility of such records or clear guidance for how such information should be integrated with other data. Part of the issue is likely the lack of uniformity regarding how records are kept, differences in the kinds of incidents or issues noted in records, and the stark differences in school or institutional settings that influence how positive and negative behavioral incidents are handled. The clinician may very well find information from records critical in validating referral concerns, but the records are essentially limited in that they likely will not include contextual information regarding the antecedents or consequences of the issue noted by the record. In that sense, records are merely descriptive and may be devoid of important contextual considerations.

Integration across informants. Recently, there have been substantive advances in the understanding of discrepancies across sources of information in child assessments (e.g., De Los Reyes, & Kazdin, 2005). This research has documented reasons for the occurrence of discrepancies, ways to reduce discrepancies that are due to error, and methods to integrate both convergent and divergent information from multiple sources. Some evidence suggests that many informant differences can be understood by referencing the situational specificity in the target child's behavior (Konold et al., 2004), yet other moderating influences on the degree of informant agreement—such as the informant's attributions, type of construct or problem, child's age or gender—also should be considered (see De Los Reyes & Kazdin, 2005).

De Los Reyes and Kazdin (2005) have proposed an attributions bias context (ABC) model that can facilitate the practitioner's recognition of general sources of informant discrepancies and can assist

decision making regarding the conduct of the assessment itself, diagnostic decisions, and the development of intervention plans from the assessment results. Clinicians should attempt to ascertain the issues most central to the child's functioning based on multiple pieces of data and must consider multiple reasons for informant discrepancies, some of which may have important implications for intervention (e.g., the child responds better to the structured environment of the school rather than the relatively unstructured home setting).

A multistep process for integrating findings across tests and informants follows from these considerations (see Frick et al., 2010). First, the clinician should document all clinically significant findings across constructs and informants. A significant finding should not be dismissed simply because of presumptions that the construct in question is irrelevant to the case or because of concerns about the accuracy of an informant's report. More information is needed before making such a decision. The collection of evidence will be critical in determining the extent to which each particular dimension of concern is part of the ultimate case conceptualization. Specifically, the clinician will still need to make further decisions regarding significant findings so that the focus is on those issues most pertinent to the child. Second, any areas in which convergence is evident across sources are noted and likely point to an area of concern (or lack of concern if there is agreement regarding the absence of symptoms on a dimension). Third, the clinician should try to determine the reasons behind any discrepancies, which may point to concerns about the informant, the test, or differences in the child's functioning across settings. The discrepancies (e.g., the child exhibits multiple behavioral problems at school but reportedly not at home) may be indicative of important issues in the case conceptualization and plans for intervention. As part of this step, the clinician should consider cultural or other systemic influences on the information obtained as well as other potential influences on the responses on assessment measures (Kazak et al., 2010; Kazdin, 2005).

For the fourth step, the clinician should develop a hierarchy of problems from primary to secondary. Secondary problems may be considered separate from the primary clinical issue or may be considered additional manifestations of the core, primary concern. For example, a child's specific academic problems may be independent of global concerns regarding inattention, or the academic problems may, in fact, be a result of inattentiveness. This hierarchy generally points to the main issue(s) that form the primary target of initial intervention efforts. Considering the extent to which a problem influences everyday functioning may provide useful guidance as to which problems are more primary versus more secondary.

In the fifth step of this process, the clinician determines the relevant information that should be in the assessment report, attempting to achieve balance between concise explanations and a clear outline of how the case conceptualization was developed and diagnostic decisions (if applicable) were made. This approach is believed to limit the chance that the clinician will unduly dismiss information that is actually important to the child's functioning while also requiring the clinician to consider the collection of converging evidence in arriving at a description of primary and secondary issues of concern.

FUTURE DIRECTIONS

In most practice settings, the professional clinician is responsible for providing services to promote successful outcomes for the child and for doing so in an ethical manner. An additional consideration is the cost effectiveness of those services. For psychological assessments, cost effectiveness translates to using sound, evidence-based, and parsimonious batteries. Further research is needed to guide clinicians more clearly on cost effectiveness and ways to convey to clients and third-party payers the validity and utility of an assessment package for the referral issue. Moreover, further development of brief assessment tools and training methods in evidence-based assessment are needed to enhance the broad appeal—and recognition of the importance—of assessment services offered in child mental health settings (Kazak et al., 2010).

Likewise, further research on newer modalities of providing assessment services is needed. Currently, there is a convergence of increased demand for child

clinical services, limited availability of assessment services in some locales, and a growth of advanced technology to permit the delivery of such services remotely. Some initial evidence suggests that psychiatric assessments by means of videoconferencing may offer a suitable alternative to more traditional face-to-face services, yet more rigorous evaluation of such novel assessment methods is necessary (see Diamond & Bloch, 2010).

In the face of calls for greater efficiency of the delivery of evidence-based services, comes recognition of factors that may lead assessments to be conducted more slowly and methodically. From the earliest point in training, clinicians should be able to recognize and appreciate the heterogeneity of presenting problems in children that then may translate to an even more detailed, comprehensive assessment. For example, child conduct problems can take on multiple forms, each with its own definition, precipitating factors, and implications for intervention (McMahon & Frick, 2005); thus, specificity in evaluating such problems is essential. This heterogeneity presents a significant challenge to efforts to provide a clear model for the assessment of a particular presenting problem or disorder (Achenbach, 2005). At the outset, the child client who is referred for an ADHD evaluation already likely does not fit perfectly with the prototype of a child with ADHD. Thus, mental health practitioners must accept and be able to address this ambiguity. Researchers should seek to develop algorithms that assist clinicians in targeting their assessment efforts more appropriately for particular disorders as well as broad referral questions. For instance, recommended approaches for assessing aggressive behavior may differ from approaches for evaluating impulsivity, even though both may start with a broader referral issue such as concerns that the child is frequently getting into trouble at school.

The discussion presented in this chapter is not intended to replace more extensive reading on current issues in evidence-based assessment, examination of literature on valid and appropriate uses of child assessment tools, and experience in conducting comprehensive psychological assessments. Instead, this overview highlights some of the current practical issues involved in child and adolescent assessments and the importance of taking a construct-centered, scientifically informed approach to assessment. For example, research on the distinction between childhood-onset versus adolescent-onset conduct problems (e.g., Moffitt, 1993) should influence the assessment of conduct problems and how clinicians conceptualize such cases. Likewise, the potential for phobias in children to be treated effectively with exposure has spurred research on the development of short-term assessment and treatment protocols for childhood phobias (e.g., Davis et al., 2009). It is hoped that the everyday conduct of psychological evaluations and the empirical literature on developmental psychopathology and assessment will continue to enjoy such a synergistic relationship where science informs assessment approaches and the practical issues that arise inform further research inquiries.

References

Achenbach, T. M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34*, 541–547. doi:10.1207/s15374424jccp3403_9

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of crossinformantcorrelations for situational specificity. *Psychological Bulletin*, *101*, 213–232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA*. Burlington: University of Vermont, Department of Psychiatry.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Barry, C. T., Frick, P. J., & Grafeman, S. J. (2008). Child versus parent report of parenting practices: Implications for the conceptualization of child behavioral and emotional problems. *Assessment, 15*, 294–303. doi:10.1177/1073191107312212

Bird, H. R., Gould, M. S., & Staghezza, B. M. (1993). Patterns of diagnostic comorbidity in a community sample of children aged 9 through 16 years. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 361–368. doi:10.1097/00004583-199303000-00018

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Minnesota Multiphasic Personality Inventory—Adolescent: Manual for administration, scoring, and*

interpretation. Minneapolis: University of Minnesota Press.

Carroll, A., Houghton, S., Taylor, M., West, J., & List-Kerz, M. L. (2006). Responses to interpersonal and physically provoking situations: The utility and application of an observation schedule for school-aged students with and without attention deficit/hyperactivity disorder. *Educational Psychology, 26*, 483–498. doi:10.1080/14616710500342424

Conners, C. K. (1995). *Conners' Continuous Performance Test*. Toronto, Ontario, Canada: Multi-Health Systems.

Conners, C. K. (2008). *Conners 3rd edition*. Toronto, Ontario, Canada: Multi-Health Systems.

Davis, T. E., Ollendick, T. H., & Ost, L. (2009). Intensive treatment of specific phobias in children and adolescents. *Cognitive and Behavioral Practice, 16*, 294–303. doi:10.1016/j.cbpra.2008.12.008

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. doi:10.1037/0033-2909.131.4.483

Diamond, J. M., & Bloch, R. M. (2010). Telepsychiatry assessments of child or adolescent behavior disorders: A review of evidence and issues. *Telemedicine and e-Health, 16*, 712–716.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development, 56*, 265–275. doi:10.2307/1130193

Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. (2005). Evidence-based assessment of learning disabilities in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 506–522. doi:10.1207/s15374424jccp3403_7

Frick, P. J., Barry, C. T., & Kamphaus, R. W. (2010). *Clinical assessment of child and adolescent personality and behavior* (3rd ed.). New York, NY: Springer. doi:10.1007/978-1-4419-0641-0

Frick, P. J., & Loney, B. R. (2000). The use of laboratory and performance-based measures in the assessment of children and adolescents with conduct disorders. *Journal of Clinical Child Psychology, 29*, 540–554. doi:10.1207/S15374424JCCP2904_7

Garb, H. N., Wood, J. M., & Lilienfeld, S. O., & Nezworski, M. T. (2005). Roots of the Rorschach controversy. *Clinical Psychology Review, 25*, 97–118. doi:10.1016/j.cpr.2004.09.002

Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment, 15*, 496–507. doi:10.1037/1040-3590.15.4.496

Kamphaus, R. W., Reynolds, C. R., & Vogel, K. K. (2009). Intelligence testing. In J. L. Maston, F. Andrasik, & M. L. Matson (Eds.), *Assessing childhood psychopathology and developmental disabilities* (pp. 91–115). New York, NY: Springer. doi:10.1007/978-0-387-09528-8_4

Kazak, A. E., Hoagwood, K., Weisz, J. R., Hood, K., Kratochwill, T. R., Vargas, L. A., & Banez, G. A. (2010). A meta-systems approach to evidence-based practice for children and adolescents. *American Psychologist, 65*, 85–97. doi:10.1037/a0017784

Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child and Adolescent Psychology, 34*, 548–558. doi:10.1207/s15374424jccp3403_10

Konold, T. R., Walthall, J. C., & Pianta, R. C. (2004). The behavior of child ratings: Measurement structure of the Child Behavior Checklist across time, informants, and child gender. *Behavioral Disorders, 29*, 372–383.

Kovacs, M. (1992). *The Children's Depression Inventory manual*. New York, NY: Multi-Health Systems.

Lahey, B. B., & Willcutt, E. G. (2010). Predictive validity of a continuous alternative to nominal subtypes of attention-deficit/hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology, 39*, 761–775. doi:10.1080/15374416.2010.517173

Lambert, M., Whipple, J., Hawkins, E., Vermeersch, D., Nielsen, S., & Smart, D. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10*, 288–301. doi:10.1093/clipsy.bpg025

Loeber, R., Green, S. M., Lahey, B. B., & Stouthamer-Loeber, M. (1991). Differences and similarities between children, mothers, and teachers as informants on childhood psychopathology. *Journal of Abnormal Child Psychology, 19*, 75–95. doi:10.1007/BF00910566

Loney, B. R., Frick, P. J., Clements, C. B., Ellis, M. L., & Kerlin, K. (2003). Callous-unemotional traits, impulsivity, and emotional processing in adolescents with antisocial behavior problems. *Journal of Clinical Child and Adolescent Psychology, 32*, 66–80.

Loney, B. R., Frick, P. J., Ellis, M., & McCoy, M. G. (1998). Intelligence, callous-unemotional traits, and antisocial behavior. *Journal of Psychopathology and Behavioral Assessment, 20*, 231–247. doi:10.1023/A:1023015318156

March, J. S., & Albano, A. M. (1996). Assessment of anxiety in children and adolescents. *American Psychiatric Press Review of Psychiatry, 15*, 405–427.

Mash, E. J., & Hunsley, J. (1997). Assessment of child and family disturbance: A developmental–systems

approach. In E. J. Mash & L. G. Terdal (Eds.), *Assessment of childhood disorders* (3rd ed., pp. 3–68). New York, NY: Guilford Press.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 362–379. doi:10.1207/s15374424jccp3403_1

Matthey, S., & Petrovski, P. (2002). The Children's Depression Inventory: Error in cutoff scores for screening purposes. *Psychological Assessment, 14*, 146–149. doi:10.1037/1040-3590.14.2.146

McConaughy, S. H., & Achenbach, T. M. (2004). *Manual for the Test Observation Form for ages 2–18*. Burlington, VT: Research Center for Children, Youth, and Families.

McMahon, R. J., & Frick, P. J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 477–505. doi:10.1207/s15374424jccp3403_6

Moffitt, T. E. (1993). Adolescence-limited and life-course persistent anti-social behavior: A developmental taxonomy. *Psychological Reports, 100*, 674–701.

O'Brien, B. S., & Frick, P. J. (1996). Reward dominance: Associations with anxiety, conduct problems, and psychopathy in children. *Journal of Abnormal Child Psychology, 24*, 223–240. doi:10.1007/BF01441486

Ollendick, T. H., & Hersen, M. (1993). *Handbook of child and adolescent assessment*. Boston, MA: Allyn & Bacon.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 523–540. doi:10.1207/s15374424jccp3403_8

Paikoff, R. L., & Brooks-Gunn, J. (1991). Do parent–child relationships change during puberty? *Psychological Bulletin*, *110*, 47–66. doi:10.1037/0033-2909.110.1.47

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention-deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 449–476.

Prinstein, M. J. (2007). Assessment of adolescents' preference- and reputation-based peer status using sociometric experts. *Merrill-Palmer Quarterly, 53*, 243–261. doi:10.1353/mpq.2007.0013

Reich, W. (2000). Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 59–66. doi:10.1097/00004583-200001000-00017

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children's Manifest Anxiety Scale (RCMAS)*. Los Angeles, CA: Western Psychological Services.

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, *112*, 485–499. doi:10.1037/0033-2909.112.3.485

Roid, G. H. (2003). *Stanford–Binet intelligence scales* (5th ed.). Itasca, IL: Riverside.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children, Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 28–38. doi:10.1097/00004583-200001000-00014

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.). Circle Pines, MN: American Guidance Service.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (4th edition): Administration and scoring manual*. San Antonio, TX: Harcourt Assessment.

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology, 15*, 403–430. doi:10.1017/S0954579403000221

CHAPTER 16

PSYCHOLOGICAL ASSESSMENT IN FORENSIC CONTEXTS

Kirk Heilbrun and Stephanie Brooks Holliday

Forensic psychology is now firmly established as a specialization in applied psychology. One of its major components is psychological assessment conducted "as part of the legal decision-making process, for the purpose of assisting the (legal) decision-maker or one of the litigants in using relevant clinical and scientific data" (Heilbrun, 2001, p. 3). For the purposes of this chapter, this kind of assessment is termed *forensic mental health assessment* (FMHA). A systems-based description of FMHA would include decisions that are made in the criminal justice system (e.g., competency to stand trial, transfer between juvenile and criminal courts, criminal responsibility, sentencing), the civil justice system (e.g., civil commitment, guardianship, personal injury, child custody), and the juvenile justice system (e.g., transfer between juvenile and criminal courts, adjudication and placement decisions; Heilbrun, Grisso, & Goldstein, 2009). Arrayed somewhat differently, FMHA tasks include description of mental states, motivations, and behaviors during past events (e.g., capacity to waive Miranda rights, criminal responsibility), deficits in abilities relevant for current contexts (e.g., guardianship and conservatorship, fitness for duty, competencies in criminal and delinquency proceedings), and predictions of future behavior and mental states (e.g., sentencing, postsentence evaluation of reoffense risk among convicted sexual offenders, child custody; Heilbrun, Douglas, & Yasuhara, 2009).

This chapter has three goals. First, we describe the relevant history and significant developments of psychological assessment in legal contexts. We focus particularly on the development of the new *Specialty Guidelines for Forensic Psychology* (Committee for Revision of the Specialty Guidelines, 2011). Second, we describe the development of specialized measures and forensically relevant tests. This distinction was made nearly a decade ago (Heilbrun, Rogers & Otto, 2002) in a review through the 1990s. We update this review to include the decade from 2000 to 2009. Although certain clinical measures (e.g., the Minnesota Multiphasic Personality Inventory) may also be used in the course of FMHA, the role of these instruments in FMHA is referenced elsewhere (for additional information, see Heilbrun et al. 2002). This chapter will focus on specialized measures developed and validated to measure constructs that are directly relevant in forensic assessment. Third, we assess the scientific support for various approaches to FMHA. In so doing, we rely particularly on two recent works—the first focusing on the available scientific literature (Skeem, Douglas, & Lilienfeld, 2009) and the second on a 20-volume series addressing best practice in FMHA.[1] For a related

[1]This Oxford University Press series includes 20 volumes on best practice in FMHA. These include the following: *Foundations of Forensic Mental Health Assessment* (Heilbrun et al., 2009), *Evaluation of Competence to Stand Trial* (Zapf & Roesch, 2009), *Evaluation of Sexually Violent Predators* (Witt & Conroy, 2008), *Evaluation of Criminal Responsibility* (Packer, 2009), *Evaluation for Risk of Violence in Adults* (Heilbrun, 2009), *Evaluation of Juveniles' Competence to Stand Trial* (Kruh & Grisso, 2009), *Evaluation for Risk of Violence in Juveniles* (Hoge & Andrews, 2009), *Evaluation of Capacity to Consent to Treatment and Research* (Kim, 2009), *Evaluation for Guardianship* (Drogin & Barrett, 2010), *Evaluation for Capital Sentencing* (Cunningham, 2010), *Evaluation of Capacity to Waive Miranda Rights* (Goldstein & Goldstein, 2010), *Jury Selection* (Kovera & Cutler, in press), *Evaluation for Harassment and Discrimination Claims* (Foote & Goodman-Delahunty, 2010), *Evaluation for Personal Injury Claims* (Kane & Dvoskin, 2011), *Evaluation for Workplace Disability* (Piechowski, 2011), *Evaluation for Civil Commitment* (Pinals & Mossman, 2011), *Evaluation for Child Protection* (Budd, Clark, Connell, & Kuehnle, 2011), and *Evaluation for Disposition and Transfer of Juvenile Offenders* (Salekin, in press).

DOI: 10.1037/14048-016
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

discussion of legal issues in assessment, readers should consult Chapter 6 in this volume.

FMHA 1960–2010: RELEVANT HISTORY

The practice of forensic assessment has developed substantially over the past 50 years. For the first 2 decades of this period, forensic assessment was not conducted very differently than evaluations performed for diagnostic and therapeutic purposes (Grisso, 1987). Such evaluations tended to use traditional measures of psychopathology, personality, academic achievement, and intellectual functioning. This was done without incorporating them into a model (such as those described by Grisso, 1986, or Morse, 1978a, 1978b) that includes both clinical characteristics and functional legal capacities. Functional legal capacities refer to what a litigant must be able to think, say, or do in order to meet a particular legal test.

Heilbrun et al. (2002) described three kinds of psychological measures that may be used in legal contexts. These include (a) clinical measures (standard psychological tests that were developed for use in diagnosis, symptom and deficit description, and intervention planning with clinical populations); (b) forensically relevant instruments (measuring clinical constructs that are sometimes pertinent to legal standards, such as psychopathy or severe depression); and (c) forensic assessment instruments (measures that are directly relevant to a specific legal standard and its included capacities that are needed for the individual being evaluated to meet that legal standard). For the most part, the psychological assessment conducted for legal purposes between 1960 and 1980 used measures of the first or second type.

There were several exceptions, however. The Checklist of Criteria for Competency to Stand Trial (Robey, 1965) was followed by other specialized measures of competence to stand trial. These included the Competency Screening Test (Lipsitt, Lelos, & McGarry, 1971), the Competency to Stand Trial Assessment Instrument (McGarry, 1971), the Georgia Court Competency Test (Wildman et al., 1979), and the Interdisciplinary Fitness Interview (Roesch & Golding, 1980).

This trend marked the beginning of an era in which there was increased attention to the development of specialized tools (see Table 16.1) as well as broader scholarship focused on the process of forensic assessment. Two of the most influential books that first appeared in the 1980s were *Evaluating Competencies* (Grisso, 1986, 2003) and *Psychological Evaluations for the Courts* (Melton, Petrila, Poythress, & Slobogin, 1987, 1997, 2007). These books were consistent with a trend involving increased attention to the research supporting the development of such specialized tools, often published in interdisciplinary journals such as *Law and Human Behavior*, *Behavioral Sciences and the Law*, and *Criminal Justice and Behavior*. As well, such research began to be published in more mainstream American Psychological Association (APA) journals such as the *Journal of Consulting and Clinical Psychology* and *Professional Psychology: Research and Practice*.

The 1980s witnessed a continuation and expansion of the development of specialized forensic assessment instruments. Some of the most significant work in this area was conducted by Grisso (1981), who used grant funding from the National Institute for Mental Health for a research project that developed four related instruments to assess the capacities of juveniles and adults to understand their Fifth and Sixth Amendment rights under *Miranda v. Arizona* (1966). This project provides both a model and a caveat for other investigators who might develop such specialized tools. It was clearly exemplary in several respects: The development and initial validation of the tool was based on carefully conducted, funded research that focused carefully on the legal demands in this context, and the properties of the measures were described in detail in the resulting book (Grisso, 1981). On the other hand, these measures were not made commercially available, with an accompanying manual describing the development, supporting research, and procedures that is so important in forensic work, until 1998 (Grisso, 1998a, 1998b). This lack undoubtedly limited their usage during the 1980s and 1990s.

The 1990s saw a very substantial increase in the number of specialized forensic assessment tools, as reflected in Table 16.1. The enhanced

TABLE 16.1

Forensic Assessment Instruments, 1960–2010

1960s	1970s	1980s	1990s	2000s
A Checklist of Criteria for Competency to Stand Trial (Robey, 1965)	Competency Assessment Instrument (McGarry, 1971)	Bricklin Perceptual Scales (Bricklin, 1984)	Ackerman–Schoendorf Parent Evaluation of Custody Test (Ackerman & Schoendorf, 1992)	Classification of Violence Risk (Monahan et al., 2006)
	Competency Screening Test (McGarry, 1971)	Custody Quotient (Gordon & Peek, 1989)	Child Abuse Potential Inventory (Milner, 1994)	Early Assessment Risk List for Boys (Augimeri et al., 2001)
	Georgia Court Competency Test (Wildman et al., 1979)	Instruments for Assessing Understanding and Appreciation of Miranda Rights (Grisso, 1981)	Competence Assessment for Standing Trial for Defendants with Mental Retardation (Everington & Luckasson, 1992)	Early Assessment Risk List for Girls (Levene et al., 2001)
		Interdisciplinary Fitness Interview (Roesch & Golding, 1980)	Computerized Assessment of Response Bias (Allen, Conder, Green, & Cox, 1992)	Estimate of Risk of Adolescent Sexual Offense Recidivism (Worling & Curwen, 2001)
		M Test (Beaber, Marston, Michelli, & Mills, 1985)	Fitness Interview Test—Revised (Roesch, Zapf, Eaves, & Webster, 1998)	Evaluation of Competency to Stand Trial—Revised (Rogers, Tillbrook, & Sewell, 2004)
		Rogers Criminal Responsibility Assessment Scales (Rogers, 1984)	HCR-20 (Webster et al., 1994)	Hare Psychopathy Checklist Revised — Youth Version (Forth, Kosson, & Hare, 2003)
			Independent Living Scales (Loeb, 1996)	Level of Service/Case Management Inventory (Andrews, Bonta & Wormith, 2004)
			Level of Service Inventory—Revised (Bonta & Andrews, 1995)	Miller Forensic Assessment of Symptoms Test (Miller, 2001)
			MacArthur Competence Adjudication Tool–Criminal Adjudication (Poythress et al., 1999)	Sex Offender Needs Assessment Rating (Hanson & Harris, 2000)
			MacArthur Competence Adjudication Tool–Treatment (Grisso & Appelbaum, 1998)	Spousal Assault Risk Assessment (Kropp & Hart, 2000)
			Malingering Probability Scale (Silverton & Gruber, 1998)	Structured Assessment for Violence Risk in Youth (Borum et al., 2006)
			Malingering Scale (Schretlen & Arkowitz, 1990)	Youth Level of Service/Case Management Inventory (Hoge & Andrews, 2002)
			Minnesota Sex Offender Screening Test (Epperson, Kaul, & Hesselton, 1998)	Violence Risk Scale (Wong & Gordon, 2001)

(*Continued*)

TABLE 16.1 (*Continued*)

Forensic Assessment Instruments, 1960–2010

1960s	1970s	1980s	1990s	2000s
			Parent Awareness Skills Survey (Bricklin, 1990b)	
			Parent Perception of Child Profile (Bricklin & Elliott, 1991)	
			Paulhus Deception Scales (Paulhus, 1998)	
			Perception-of-Relationships Test (Bricklin, 1990a)	
			Psychopathy Checklist—Revised (Hare, 1991)	
			Rapid Risk Assessment for Sex Offender Recidivism (Hanson, 1997)	
			Sex Offender Risk Appraisal Guide (Quinsey et al., 1998)	
			Sexual Violence Recidivism–20 (Boer et al., 1997)	
			Spousal Assault Risk Assessment (Kropp et al., 1995)	
			Static-99 (Hanson & Thornton, 2000)	
			Structured Inventory of Malingered Symptoms (Smith, 2002)	
			Structured Interview of Reported Symptoms (Rogers et al., 1992)	
			Test of Memory Malingering (Tombaugh, 1996)	
			Uniform Child Custody Evaluation System (Munsinger & Karlson, 1994)	
			Validity Indicator Profile (Frederick, 1997)	
			Victoria Symptom Validity Test (Slick et al., 1997)	
			Violence Prediction Scheme (Webster et al., 1994)	
			Violence Risk Appraisal Guide (Quinsey et al., 1998)	

Note. Adapted from *Taking Psychology and Law Into the Twenty-First Century* (pp. 124–125), by J. R. P. Ogloff (Ed.), 2002, New York, NY: Kluwer Academic/Plenum.

demand for such tools was apparently driven by several influences. The field of forensic psychology was maturing, with a commensurate increase in the professional and scientific literature, the coherence of the field as a whole, the availability of specialized training at multiple levels (Melton et al., 1997). Because of the economic and professional impact of managed care on the practice of applied psychology, there was more interest in forensic psychology on the part of clinical psychologists who had not specialized in the area (Otto, 1999). Consistent with both the maturation of the field and the changing context of psychological practice, publishers of psychological tests and books turned to this field as fertile ground for expanding their own products (Otto, 1999).

Understanding the development of forensic psychology, particularly that aspect pertaining to forensic mental health assessment, is facilitated by reviewing the *Specialty Guidelines for Forensic Psychologists* (Committee on Ethical Guidelines for Forensic Psychologists, 1991) and its revision, the *Specialty Guidelines for Forensic Psychology* (Committee for Revision of the Specialty Guidelines, 2011). The 1991 *Specialty Guidelines* represent the first effort by the field of forensic psychology to provide ethical guidelines that were more specific than those offered for the broader field of psychology (at the time, the *Ethical Principles of Psychologists and Code of Conduct*; American Psychological Association, 1990). Notably, the *Specialty Guidelines* were written to describe a domain encompassing a broad range of professional services provided by psychologists in legal contexts. Their applicability was described as follows:

> The Guidelines provide an aspirational model of desirable professional practice by psychologists, within any subdiscipline of psychology (e.g., clinical, developmental, social, experimental), when they are engaged regularly as experts and represent themselves as such, in an activity primarily intended to provide professional psychological expertise to the judicial system. (Committee on Ethical Guidelines for Forensic Psychologists, 1991, p. 656)

There are several particularly important points regarding the 1991 *Specialty Guidelines*. First, although their applicability was broadly defined, they clearly encompassed psychological evaluations performed for courts and attorneys. Second, they contained a substantial exception: applying to forensic psychologists who "provide services only in areas of psychology in which they have specialized knowledge, skill, experience, and education" (Committee on Ethical Guidelines for Forensic Psychologists, 1991, p. 658), they nevertheless excluded those who were not "engaged regularly as experts and represent(ed) themselves as such" (Committee on Ethical Guidelines for Forensic Psychologists, 1991, p. 656). Third, they described differences in documentation and decision making that were at the heart of how forensic assessment differs from psychological assessment conducted for diagnostic and treatment-planning purposes. Documentation demands were expected to be greater in forensic assessment. Such requirements included both the expectation that raw data (interview notes for the litigant and collateral observers, psychological test data, record reviews) would be carefully collected and made available for review by opposing counsel, and that the relationship between data and conclusions be clear from the description of the forensic clinician's reasoning. Finally, the 1991 *Specialty Guidelines* made it clear that they were part of a broader array of ethical standards and guidelines, referencing the APA *Ethical Principles* as enforceable and other documents (e.g., *Standards for Educational and Psychological Testing*; American Educational Research Association, APA, & National Council on Measurement in Education, 1999) as aspirational but also applicable.

The revision of the 1991 *Specialty Guidelines* has been ongoing since 2002. In part, this effort has resulted from the decision to seek broader approval within the field of psychology. The 1991 *Specialty Guidelines* were a joint project of APA Division 41 (American Psychology–Law Society [AP-LS]) and the American Academy of Forensic Psychology. Approved by both organizations, they were published in *Law and Human Behavior* (the official journal of AP-LS). APA no longer permits divisions to

promulgate ethical guidelines independent of APA review, however, so the *Specialty Guidelines*' revision has undergone periods of review by other divisions and committees within APA and will eventually require approval by the APA Council of Representatives, the organization's governance body, before they are adopted.

Although the revised *Specialty Guidelines* affirms many of the points made in the original document, there have been several noteworthy and important changes. First, the revised version is titled *Specialty Guidelines for Forensic Psychology,* with the nature of forensic psychology described as

> professional practice by any psychologist working within any sub-discipline of psychology (e.g., clinical, developmental, social, cognitive) when applying the scientific, technical, or specialized knowledge of psychology to the law to assist in addressing legal, contractual, and administrative matters. (Committee for Revision of the Specialty Guidelines, 2011)

In contrast to the 1991 *Specialty Guidelines*, the revised *Guidelines* clearly reflect the intended relevance to all forensic psychological activities and focus on the work (however infrequently performed) rather than the individuals who provide it.

The other major distinction between the two documents involves expanded sections on "Notification, Assent, Consent, and Informed Consent" and "Assessment," the latter devoted to the nature of forensic assessment itself. These sections reflect a considerable expansion in the literature addressing these topics (e.g., see Heilbrun, 2001; Melton et al., 1997, 2007) during the interim between publication of the original *Guidelines* and the current revision.

SPECIALIZED FORENSIC ASSESSMENT MEASURES: 2000–2010

If the decade from 1990 to 1999 could be fairly described as witnessing a virtual explosion in forensic specialty tools, particularly in those assessing response style, then the subsequent 10 years might be called "the decade of risk assessment." Stimulated by the MacArthur study of mental disorder and violence (Monahan et al., 2001), researchers conducted work devoted to the development, validation, and refinement of various specialized risk assessment tools at a brisk pace. Indeed, of the 13 measures described in Table 16.1 that were developed or revised from 2000 to 2009, one is a measure of the capacities associated with competence to stand trial (the Evaluation of Competency to Stand Trial—Revised; Rogers, Tillbrook, & Sewell, 2004) and two others are forensically relevant tests (the Hare Psychopathy Checklist Revised—Youth Version; Forth, Kosson, & Hare, 2003; and the Miller Forensic Assessment of Symptoms Test; Miller, 2001). Those remaining are all specialized risk assessment measures.

There is considerable variability within this group of risk assessment measures. The Level of Service/Case Management Inventory (Andrews, Bonta, & Wormith, 2004) is the revision of the Level of Service Inventory (Andrews, 1982) and Level of Service Inventory—Revised (Andrews & Bonta, 1997), both risk-needs tools developed for assessing risk of reoffending and risk-relevant deficits in general correctional populations, based on risk-need-responsivity theory (Andrews, Bonta, & Hoge, 1990). This kind of risk-needs assessment, in which both static and dynamic risk factors are included in an appraisal of overall risk of reoffense or violence *and* the identification of risk-relevant intervention targets, appeared to be particularly useful in assessing the risk of adolescents involved in the juvenile justice system. The first measure of adolescent risk described in Table 16.1 is the Youth Level of Service/Case Management Inventory (Hoge & Andrews, 2002), an adaptation of the adult version. The second is the Structured Assessment for Violence Risk in Youth (Borum, Bartel, & Forth, 2006). These measures are similar in a number of respects, both encompassing empirically supported domains of risk factors for adolescents (see Andrews & Hoge, 2010) and forcing the evaluator to attend to the adolescent's functioning in each of these domains in order to reach a final risk estimate and describe relevant intervention needs. A similar approach was taken in the development of an early risk assessment measure for boys and girls (Augimeri, Webster, Koegl, & Levene, 2001; Levene et al., 2001),

although these "early" measures are much less likely to be a part of juvenile proceedings because the individuals described are younger than those typically subject to juvenile jurisdiction.

The Classification of Violence Risk (Monahan et al., 2006) is noteworthy for two reasons in the context of this discussion. It is a measure developed to estimate the risk of individuals with severe mental illness for violence in the community. As such, it differs from all the other adult risk assessment measures in Table 16.1 in its exclusive applicability to individuals who are not involved in the criminal justice system. It is also the best example of a strong actuarial risk assessment measure, originally developed from the MacArthur risk data set (Monahan et al., 2001) and validated in another two-site study (Monahan et al., 2005). It represents the best kind of actuarial tool available: developed and validated using large data sets, manualized with clear directions for administration and interpretation, and accompanied by additional information describing narrow, nonoverlapping 95% confidence intervals surrounding each risk category and guiding the user in communicating the resulting risk estimate. The debate between using risk assessment tools that are actuarial versus those using structured professional judgment (see Heilbrun, Douglas, et al., 2009, for a summary of this debate) has included critical comments regarding actuarial measures that are not developed and validated with sufficiently large samples, and/or yielding risk categories that overlap—making it more difficult to determine accurately whether an individual is appropriately placed in one risk category or the other.

In this context, there were three additional noteworthy developments in risk assessment during 2000–2009. Along with updating the Static-99 into the Static-2002 (Hanson & Thornton, 2003), there has been additional related work on the dynamic (changeable through planned intervention) risk factors for sexual offenders. Anderson and Hanson (2009) have developed the Sex Offender Needs Assessment Rating, which includes both stable (Stable-2000, Stable-2007) and acute (Acute-2000, Acute-2007) dynamic risk factors. This development is an important step for researchers who develop actuarial tools that rely exclusively on static risk factors, primarily from an individual's history. The legal demand for risk reduction appears as strong as the demand for classification of risk level; accordingly, tools that can provide accurate information in both domains offer some advantage over actuarial measures that provide only risk classification guidance. In this vein, the second positive development in risk assessment during 2000–2009 has involved the development of the Violence Risk Scale (Wong & Gordon, 2001). This scale represents a somewhat different approach to combining actuarial risk assessment with both static and dynamic risk factors. The VRS has a solid theoretical foundation, based on risk, need and responsivity principles. It is intended for use by scientists/practitioners to assess and predict the risk of violence, to measure changes in risk after treatment, and to make treatment decisions. This approach to combining both static and dynamic risk factors in actuarial fashion, also seen in the Level of Service Inventory measures, involves a very promising use of several aspects of risk assessment for multiple purposes.

The final interesting and positive development of the decade involves the Risk for Sexual Violence Protocol (RSVP; Hart & Boer, 2009). As a relatively new tool, the RSVP probably needs additional research before it is used in practice. It is a structured professional judgment tool, conceptualizing risk to include nature, severity, imminence, frequency, and likelihood of sexual offending. Initial research reflects good to excellent reliability of the 22 items over five domains (sexual violence history, psychosocial adjustment, mental disorder, social adjustment, and manageability). With additional research on its validity in different populations, it has the potential to become a primary specialized tool in the assessment of sexual offending risk.

THE CURRENT SCIENTIFIC FOUNDATIONS OF FMHA: IMPLICATIONS FOR BEST PRACTICE

One of the recent valuable additions to the literature relating scientific evidence to FMHA is an edited volume titled *Psychological Science in the Courtroom: Consensus and Controversy* (Skeem, Douglas, &

Lilienfeld, 2009). Authors contributing chapters were asked to describe the foundational science relevant to various legal questions, and summarize the "scientifically supported," "scientifically unsupported," and "scientifically controversial" activities pursued in the course of forensic assessment of different kinds. It is useful to summarize these conclusions drawn by authors addressing the assessment of psychological injuries (Koch, Nader, & Haring, 2009), child custody evaluations (O'Donohue, Beitz, & Tolle, 2009), competence to stand trial (Poythress & Zapf, 2009), and violence risk assessment (Heilbrun, Douglas, et al., 2009), respectively.

In the domain involving the forensic assessment of psychological injury (Koch et al., 2009), the following were described as "scientifically supported" uses:

- Psychological evaluation of current mental health functioning, as long as it includes cautions regarding response style in litigating samples and the limits of current assessment of such response style;
- Descriptions of reports of past functioning provided by claimants and collateral interviewees, again accompanied by caveats pertaining to limits of this task; and
- Use of well-validated measures of symptom overendorsement, applied to opinions about claimants' response style during the present assessment and not necessarily to a larger conclusion regarding malingering.

One scientifically untested or controversial use was noted—conclusions regarding the effect of psychological injury on particular areas of future disability (e.g., capacity to perform some but not other work tasks; precise estimates about the duration of functional incapacity in certain domains). Finally, several "scientifically unsupported" uses were described:

- Definitive conclusions that current incapacity was caused by the legally contested adverse event;
- Definitive descriptions of past psychological functioning derived from current psychological testing;
- Conclusions that an individual is malingering without strong evidence of intentionality; and
- Prognoses about future psychological functioning or disability made using mental health variables alone.

In the area of child custody evaluations (O'Donohue et al., 2009), the authors offered two conclusions regarding scientifically supported procedures:

- Conclusions about current psychological or psychiatric functioning of children, including special needs, mental health functioning, and problems that can be addressed through planned intervention; and
- Research evidence on the impact of divorce on children.

Scientifically untested or controversial uses included predictions about children's functioning in the future using tests or measures with sound psychometric properties for present assessment. The major scientifically unsupported use was described as predictions about custody arrangements in the best interests of children using specialized measures that have not been validated for making such predictions.

The next chapter in Skeem, Douglas, and Lilienfeld (2009) concerned the evaluation of competence to stand trial (Poythress & Zapf, 2009). Scientifically supported approaches to such evaluations were described as:

- Incorporating the results of idiographic assessment measures (those assessing the individual's functioning relative to his/her own potential and prior performance);
- Incorporating results of nomothetic assessment measures (assessing the individual's functioning relative to known groups); and
- Using either or both in describing the competence-related abilities of the individual being evaluated.

The single scientifically untested or controversial use was noted to involve relying solely on data from idiographic assessment (without incorporating nomothetic assessment results) in drawing conclusions regarding the defendant's competence-relevant

capacities. By contrast, the scientifically unsupported uses of assessment approaches in this area were given as:

- Combining clinical ratings from different idiographic items or "scales" and interpreting these scores as meaningfully relevant to trial competence;
- Using nomothetic measures, without case-specific idiographic inquiry, as the only basis for conclusions relevant to trial competence; and
- Combining scores on nomothetic measures toward a dichotomous conclusion regarding trial competence.

In addition, Heilbrun, Douglas, et al. (2009) addressed violence risk assessment, a procedure that can be used to inform the court in making decisions regarding a number of legal questions (e.g., civil commitment, criminal sentencing, commitment and release of individuals in categories such as Violent Sexual Predators or Not Guilty by Reason of Insanity acquittees). Scientifically supported uses of risk assessment procedures were described as follows:

- Conclusions that those scoring higher on validated risk assessment instruments are at greater risk for violence;
- Actuarial prediction strategies applied to groups, obtained using large derivation and validation samples, citing mean probability, and specifying margin of error;
- Use of extreme risk categories as both more informative and less affected by the possibility of overlapping 95% confidence intervals; and
- Indication that applying group-based data to a small number of cases (or an individual case) will result in wider confidence intervals than when such data are applied to a larger number of cases.

Two scientifically untested or controversial uses were cited:

- Actuarial prediction strategies with appropriately large derivation and validation samples and correctly citing mean probability but not specifying margin of error and its increased uncertainty when applied to single cases; and
- Assuming the existence of reliable, known probability estimates (robust across samples) even at the group level.

Finally, the authors described three scientifically unsupported uses of risk assessment:

- Actuarial prediction strategies that were developed without the requisite large derivation and validation samples;
- Applying actuarial prediction approaches to populations without empirical foundation (via either the derivation or validation samples) for doing so; and
- Drawing conclusions about the probability of an individual's future violence without providing applicable confidence intervals associated with the prediction measure and caution about less certainty in the individual case.

What broad conclusions regarding the scientific foundations of FMHA can be drawn from these reviews? There appear to be several, related both to the availability of foundational scientific data and to the conceptualization, data-gathering, interpretation, and reasoning process in FMHA that is modeled on scientific thinking. First, it is clear that a number of authors have emphasized the importance of specialized forensic assessment measures. Such measures directly link the measurement of relevant legal capacities with the supporting scientific evidence. When there is a specialized measure that has been appropriately derived and validated, it should be used as part of the assessment of an individual in a given case.

The second conclusion concerns the importance of multiple sources of information in FMHA. Consistent with the measurement of human functioning in multitrait, multimethod fashion (Campbell & Fiske, 1959), these authors have underscored the importance of obtaining information both from the individual being evaluated (and using both nomothetic and idiographic measures) and from collateral sources such as third party interviews and relevant records. This posture relates directly to a third conclusion regarding the scientific foundations of FMHA: the importance of avoiding "overreaching." Drawing conclusions in the course of individual

evaluations without adequate support could be done in different ways. The evaluator could attempt to use a measure in a way that is beyond how it was intended or validated for use; for example, using psychological tests to predict future functioning rather than describing the present state. Applying a single measure toward drawing a conclusion when the results of that measure are not supported by information from other sources represents another form of overreaching. A third example might involve reconfiguring or combining existing measures without empirical support for doing so.

The final conclusion that may be drawn from these reviews concerning the relationship between science and FMHA concerns the appropriate use of cautions. Measures are often limited in their applicable scope. Data may be inconsistent or missing. Some conclusions can be reached with confidence, given the nature of the data collected; others must be more tentative (or, in some cases, cannot be drawn). The explicit description of the limitations of the data, and the impact of such limitations on the conclusions that can be drawn, are an important philosophical application of scientific methodology to FMHA.

The Oxford Series ("Best Practices in Forensic Mental Health Assessment" described in Footnote 1) involves an effort to characterize both the foundations of FMHA (Heilbrun, Grisso, et al., 2009) and the more specific assessment procedures considered by legal questions. This task probably could not have been accomplished even a decade ago, reflecting the kind of progress that has been witnessed during this time. Nevertheless, there is inconsistency observed in the advances made in different areas. Series authors made a concerted effort to incorporate scientific, ethical, legal, and professional sources of authority into their conclusions regarding best practice. However, some areas have received relatively less research attention without the development of a specialized forensic assessment measure (e.g., juvenile competence to stand trial), whereas others (e.g., risk assessment) have been the subject of intensive research and witnessed the development of multiple specialized measures. Continued scientific attention, hopefully distributed in a manner that enhances some of these less developed areas, will be important in the coming decade. It also would be helpful if descriptions of best practice across various legal questions were addressed through practice guidelines. Such guidelines would be developed by a major organization representing the field, such as the AP–LS, rather than through the efforts of three editors and specific authors in a book series. In that respect, the discipline-endorsed practice guidelines would have more credibility as representing an aspirational standard of practice in the field.

CONCLUSION

There have been a number of important developments in forensic mental health assessment during the near-decade since the Heilbrun et al. (2002) summary. Such advances include the development of additional specialized forensic assessment instruments, and an even clearer recognition of their value. This period has also included a number of efforts, cited in this chapter, to characterize the state of the field and describe its strengths and needs. Among the most pressing current needs are the development of specialized forensic measures in areas in which they do not currently exist and the description of practice guidelines by a larger organization, representative of forensic psychology and having the potential to provide discipline-level endorsement.

References

Ackerman, M., & Schoendorf, K. (1992). *The Ackerman-Schoendorf parent evaluation of custody tests (ASPECT)*. Los Angeles, CA: Western Psychological Services.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (1990). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author.

Anderson, D., & Hanson, K. (2009). An actuarial tool to assess risk of sexual and violent recidivism among sexual offenders. In R. Otto & K. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 251–267). New York, NY: Routledge.

Andrews, D. (1982). *The Level of Supervision Inventory (LSI): The first follow-up.* Toronto, Ontario, Canada: Ministry of Correctional Services.

Andrews, D., & Bonta, J. (1997). *Manual for the Level of Service Inventory—Revised.* North Tonawanda, NY: Multi-Health Systems.

Andrews, D., Bonta, J., & Hoge, R. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*, 19–52. doi:10.1177/0093854890017001004

Andrews, D., Bonta, J., & Wormith, J. (2004). *Level of service/case management inventory (LS/CMI).* Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., & Hoge, R. (2010). *Evaluation for risk of violence in juveniles.* New York, NY: Oxford University Press.

Augimeri, L., Webster, C., Koegl, C., & Levene, K. (2001). *Early Assessment Risk List for Boys (EARL-20B): Version 2.* Toronto, Ontario, Canada: Earlscourt Child and Family Centre.

Beaber, J. R., Marston, A., Michelli, J., & Mills, M. J. (1985). A brief test for measuring malingering in schizophrenic individuals. *American Journal of Psychiatry, 142*, 1478–1481.

Boer, D., Hart, S., Kropp, P., & Webster, C. (1997). *Manual for the Sexual Violence Risk-20.* Burnaby, British Columbia, Canada: Mental Health, Law, & Policy Institute, Simon Fraser University.

Bonta, J., & Andrews, J. (1995). *Manual for the Level of Service Inventory—Revised.* Tonawanda, NY: Multi-Health Systems.

Borum, R. Bartel, P., & Forth, A. (2006). *Manual for the Structured Assessment of Violence Risk in Youth (SAVRY).* Odessa, FL: Psychological Assessment Resources.

Bricklin, B. (1984). *Bricklin Perceptual Scales manual.* Furlong, PA: Village.

Bricklin, B. (1990a). *Perceptions of Relationships Test manual.* Furlong, PA: Village.

Bricklin, B. (1990b). *Parent Awareness Skills Survey manual.* Furlong, PA: Village.

Bricklin, B., & Elliott, G. (1991). *Parent Perception of Child Profile Manual.* Furlong, PA: Village Publishing.

Budd, K., Clark, J., Connell, M., & Kuehnle, K. (2011). *Evaluation for child protection.* New York, NY: Oxford University Press.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty guidelines for forensic psychologists. *Law and Human Behavior, 15*, 655–665. doi:10.1007/BF01065858

Committee for Revision of the Specialty Guidelines. (2011). *Specialty guidelines for forensic psychology.* Retrieved from http://www.ap-ls.org/aboutpsychlaw/SpecialtyGuidelines.php

Cunningham, M. (2010). *Evaluations for capital sentencing.* New York, NY: Oxford University Press.

Drogin, E., & Barrett, C. (2010). *Evaluation for guardianship.* New York, NY: Oxford University Press.

Epperson, D., Kaul, J., & Hesselton, D. (1998, September). *Final report on the development of the Minnesota Sex Offender Screening Tool—Revised (MnSOST-R).* Presented at the annual meeting of the Association for the Treatment of Sexual Abusers, Vancouver, British Columbia, Canada.

Everington, C., & Luckasson, R. (1992). *Manual for Competence Assessment for Standing Trial for Defendants with Mental Retardation: CAST-MR.* Worthington, OH: IDS Publishing.

Foote, W., & Goodman-Delahunty, J. (2010). *Evaluation for harassment and discrimination claims.* New York, NY: Oxford University Press.

Forth, A., Kosson, D., & Hare, R. (2003). *Hare Psychopathy Checklist: Youth Version (PCL:YV).* North Tonawanda, NY: Multi-Health Systems.

Frederick, R. (1997). *Manual for the Validity Indicator Profile.* Minnetonka, MN: National Computer Services.

Goldstein, A., & Goldstein, N. (2010). *Evaluating capacity to waive Miranda rights.* New York, NY: Oxford University Press.

Gordon, R., & Peek, L. (1989). *The Custody Quotient: Research manual.* Dallas, TX: Wilmington Institute.

Grisso, T. (1981). *Juveniles' waiver of rights: Legal and psychological competence.* New York, NY: Plenum Press.

Grisso, T. (1986). *Evaluating competencies: Forensic assessments and instruments.* New York, NY: Plenum Press.

Grisso, T. (1987). The economic and scientific future of forensic psychological assessment. *American Psychologist, 42*, 831–839. doi:10.1037/0003-066X.42.9.831

Grisso, T. (1998a). *Instruments for assessing understanding and appreciation of Miranda rights.* Sarasota, FL: Professional Resource Press.

Grisso, T. (1998b). *Instruments for assessing understanding and appreciation of Miranda rights—Manual.* Sarasota, FL: Professional Resource Press.

Grisso, T. (2003). *Evaluating competencies: Forensic assessments and instruments* (2nd ed.). New York, NY: Springer Science + Business Media.

Grisso, T., & Appelbaum, P. (1998). *MacArthur Competence Assessment Tool–Treatment (MacCAT-T)*. Sarasota, FL: Professional Resource Press.

Hanson, K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism* (User Report No. 1997–04). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, K., & Harris, A. (2000). *The sex offender needs assessment rating (SONAR): A method of measuring change in risk levels.* Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24*, 119–136. doi:10.1023/A:1005482921333

Hanson, K., & Thornton, D. (2003). *Notes on the development of the Static-2002.* Ottawa, Ontario, Canada: Solicitor General of Canada (Corrections Research User Report 2003–01).

Hare, R. (1991). *Manual for the Hare Psychopathy Checklist—Revised.* North Tonawanda, NY: Multi-Health Systems.

Hart, S. D., & Boer, D. P. (2009). Structured Professional Judgment guidelines for sexual violence risk assessment: The Sexual Violence Risk (SVR-2) and Risk for Sexual Violence Protocol (RSVP). In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (pp. 269–294). New York, NY: Routledge.

Heilbrun, K. (2001). *Principles of forensic mental health assessment.* New York, NY: Kluwer Academic/Plenum Press.

Heilbrun, K. (2009). *Evaluation for risk of violence in adults.* New York, NY: Oxford University Press.

Heilbrun, K., Douglas, K., & Yasuhara, K. (2009). Violence risk assessment: Core controversies. In J. Skeem, K. Douglas, & S. Lilienfeld (Eds.), *Psychological science in the courtroom: Controversies and consensus* (pp. 333–357). New York, NY: Guilford Press.

Heilbrun, K., Grisso, T., & Goldstein, A. (2009). *Foundations of forensic mental health assessment.* New York, NY: Oxford University Press.

Heilbrun, K., Rogers, R., & Otto, R. K. (2002). Forensic assessment: Current status and future directions. In J. R. P. Ogloff (Ed.), *Taking psychology and law into the twenty-first century* (pp. 119–146). New York, NY: Kluwer Academic/Plenum.

Hoge, R., & Andrews, D. (2002). *Youth Level of Service/Case Management Inventory (YLS/CMI).* Toronto, Ontario, Canada: Multi-Health Systems.

Hoge, R., & Andrews, D. (2009). *Evaluation of risk of violence in juveniles.* New York, NY: Oxford University Press.

Kane, A., & Dvoskin, J. (2011). *Evaluation for personal injury claims.* New York, NY: Oxford University Press.

Kim, S. (2009). *Evaluation of capacity to consent to treatment and research.* New York, NY: Oxford University Press.

Koch, W., Nader, R., & Haring, M. (2009). The science and pseudoscience of assessing psychological injuries. In J. Skeem, K. Douglas, & S. Lilienfeld (Eds.), *Psychological science in the courtroom: Consensus and controversy* (pp. 263–283). New York, NY: Guilford Press.

Kovera, M., & Cutler, B. (in press) *Jury selection.* New York, NY: Oxford University Press.

Kropp, P., Hart, S., Webster, C., & Eaves, D. (1995). *Manual for the Spousal Assault Risk Assessment Guide.* Vancouver, Canada: British Columbia Institute on Family Violence.

Kropp, R., & Hart, S. (2000). The Spousal Assault Risk Assessment Guide (SARA): Reliability and validity in adult male offenders. *Law and Human Behavior, 24*, 101–118. doi:10.1023/A:1005430904495

Kruh, I., & Grisso, T. (2009). *Evaluation of juveniles' competence to stand trial.* New York, NY: Oxford University Press.

Levene, K., Augimeri, L., Pepler, D., Walsh, M., Koegle, C., & Webster, C. (2001). *Early Assessment Risk List for Girls: EARL-21G, Version 1, Consultation edition.* Toronto, Ontario, Canada: Earlscourt Child and Family Centre.

Lipsitt, P. D., Lelos, D., & McGarry, A. (1971). Competency for trial: A screening instrument. *American Journal of Psychiatry, 128*, 105–109.

Loeb, P. (1996). *Manual for the Independent Living Scales.* San Antonio, TX: Psychological Corporation.

McGarry, A. (1971). *Competency to stand trial and mental illness.* Rockville, MD: U.S. Department of Health, Education and Welfare.

Melton, G., Petrila, J., Poythress, N., & Slobogin, C. (1987). *Psychological evaluations for the courts: A handbook for attorneys and mental health professionals.* New York, NY: Guilford Press.

Melton, G., Petrila, J., Poythress, N., & Slobogin, C. (1997). *Psychological evaluations for the courts: A handbook for attorneys and mental health professionals* (2nd ed.). New York, NY: Guilford Press.

Melton, G., Petrila, J., Poythress, N., & Slobogin, C. (2007). *Psychological evaluations for the courts: A handbook for attorneys and mental health professionals* (3rd ed.). New York, NY: Guilford Press.

Miller, H. (2001). *Miller Forensic Assessment of Symptoms Test (M-FAST).* Lutz, FL: Psychological Assessment Resources.

Milner, J. (1994). Assessing physical child abuse risk: The Child Abuse Potential Inventory. *Clinical Psychology Review, 14*, 547–583. doi:10.1016/0272-7358(94)90017-5

Miranda v. Arizona, 384 U.S. 436 (1966).

Monahan, J., Steadman, H., Appelbaum, P., Grisso, T., Mulvey, E., Roth, L., . . . Silver, E. (2006). The classification of violence risk. *Behavioral Sciences and the Law*, *24*, 721–730. doi:10.1002/bsl.725

Monahan, J., Steadman, H., Robbins, P., Appelbaum, P., Banks, S., Grisso, T., . . . Silver, E. (2005). Prospective validation of the multiple iterative classification tree model of violence risk assessment. *Psychiatric Services*, *56*, 810–815. doi:10.1176/appi.ps.56.7.810

Monahan, J., Steadman, H., Silver, E., Appelbaum, P., Robbins, P., Mulvey, E., . . . Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York, NY: Oxford University Press.

Morse, S. (1978a). Crazy behavior, morals, and science: An analysis of mental health law. *Southern California Law Review, 51*, 527–654.

Morse, S. (1978b). Law and mental health professionals: The limits of expertise. *Professional Psychology, 9*, 389–399. doi:10.1037/0735-7028.9.3.389

Munsinger, H. L., & Karlson, K. W. (2004). *Uniform Child Custody Evaluation System (UCCES)*. Odessa, FL: Psychological Assessment Resources.

O'Donohue, W., Beitz, K., & Tolle, L. (2009). Controversies in child custody evaluations. In J. Skeem, K. Douglas, & S. Lilienfeld (Eds.), *Psychological science in the courtroom: Consensus and controversy* (pp. 284–308). New York, NY: Guilford Press.

Otto, R. (1999, February). *The future of forensic psychology: A view towards the future in light of the past*. Paper presented at Sam Houston State University, Department of Psychology and College of Criminal Justice.

Packer, I. (2009). *Evaluation of criminal responsibility*. New York, NY: Oxford University Press.

Paulhus, D. L. (1998). *Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding—7*. North Tonawanda, NY: Multi-Health Systems.

Piechowski, L. (2011). *Evaluation for workplace disability*. New York, NY: Oxford University Press.

Pinals, D., & Mossman, D. (2011). *Evaluation for civil commitment*. New York, NY: Oxford University Press.

Poythress, N., Monahan, J., Bonnie, R., & Hoge, S. (1999). *MacArthur competence assessment tool—Criminal adjudication*. Odessa, FL: Psychological Assessment Resources.

Poythress, N., & Zapf, P. (2009). Controversies in evaluating competence to stand trial. In J. Skeem, K. Douglas, & S. Lilienfeld (Eds.), *Psychological science in the courtroom: Consensus and controversy* (pp. 309–329). New York, NY: Guilford Press.

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association. doi:10.1037/10304-000

Robey, A. (1965). Criteria for competency to stand trial: A checklist for psychiatrists. *American Journal of Psychiatry, 122*, 616–623.

Roesch, R., & Golding, S. (1980). *Competency to stand trial*. Urbana-Champaign: University of Illinois Press.

Rogers, R. (1984). *Manual for the Rogers Criminal Responsibility Assessment Scales*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., Bagby, M., & Dickens, S. (1992). *Manual for the Structured Interview of Reported Symptoms*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., Tillbrook, C., & Sewell, K. (2004). *Evaluation of competency to stand trial—Revised (ECST-R) and professional manual*. Lutz, FL: Psychological Assessment Resources.

Salekin, R. (in press). *Evaluation for disposition and transfer of juvenile offenders*. New York, NY: Oxford University Press.

Schretlen, D., & Arkowitz, H. (1990). A psychological test battery to detect prison inmates who fake insanity or mental retardation. *Behavioral Sciences and the Law, 8*, 75–84. doi:10.1002/bsl.2370080109

Silverton, L., & Gruber, C. (1998). *Manual for the Malingering Probability Scale*. Los Angeles, CA: Western Psychological Services.

Skeem, J., Douglas, K., & Lilienfeld, S. (Eds.). (2009). *Psychological science in the courtroom: Controversies and consensus*. New York, NY: Guilford Press.

Slick, D., Hopp, G., Strauss, E., & Thompson, G. (1997). *Professional manual for the Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources.

Smith, G. P. (2002). *Structured Inventory of Malingered Symptomotology*. Odessa, FL: Psychological Assessment Resources.

Tombaugh, T. (1996). *Manual for the Test of Memory Malingering*. North Tonawanda, NY: Multi-Health Systems.

Webster, C., Harris, G., Rice, M., Cormier, C., & Quinsey, V. (1994). *The Violence Prediction Scheme: Assessing dangerousness in high risk men*. Toronto, Ontario, Canada: Centre of Criminology, University of Toronto.

Wildman, R., Batchelor, E., Thompson, L., Nelson, F., Moore, J., Patterson, M., & deLaosa, M. (1979). *The Georgia court competency test*. Unpublished

manuscript, Forensic Services Division, Central State Hospital, Milledgeville, GA.

Witt, P., & Conroy, M. A. (2008). *Evaluation of sexually violent predators.* New York, NY: Oxford University Press.

Wong, S., & Gordon, A. (2001). The Violence Risk Scale. *Forensic Update, 67*, 19–23.

Worling, J., & Curwen, T. (2001). Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR), Version 2.0. In M. C. Calder (Ed.), *Juveniles and children who sexually abuse: Frameworks for assessment* (pp. 372–397). Dorset, United Kingdom: Russell House.

Zapf, P., & Roesch, R. (2009). *Evaluation of competence to stand trial.* New York, NY: Oxford University Press.

CHAPTER 17

PSYCHOLOGICAL ASSESSMENT IN MEDICAL SETTINGS

Elizabeth M. Altmaier and Benjamin A. Tallman

One of the most exciting places for psychologists to practice is in health care settings. This excitement comes about as health care is increasingly a context that prompts innovation in practice and policy development. One of the most compelling arguments for psychological practice in health care settings is the growing emphasis nationally on the integration of psychology into primary care. As described by Gray, Brody, and Johnson (2005), primary care is "behavioral health management." Thus, a comprehensive system with psychologists who are equal partners in a multidisciplinary health care team has been recognized as a reform that can meet the critical goals of health care cost containment as well as demonstrate improved patient outcomes. In fact, James and Folen (2005) described primary care as psychology's "next frontier."

Psychologists are not new arrivals to the health care setting. Belar and Deardorff (2009) traced the development of clinical health psychology back to the 1970s: They noted that psychologists initially were involved primarily as researchers, investigating psychological predictors or sequelae of medical disorders. With regard to psychological practice, early concerns centered on how psychology, as a discipline concerned with mental health, related to disorders of physical health. (For a review of the interaction of organized psychology with the Joint Commission on Accreditation of Hospitals, see Zaro, Batchelor, Ginsberg, & Pallak, 1982.) Also, psychology's longstanding adherence to a mind–body dualism that locates client issues in either the physical or mental domain, with no intersection of these domains, worked against psychologists obtaining a practice role in health care. However, the contributions of clinical health psychologists in research and in practice were undeniable; and the domains of assessment, intervention, and consultation have continued to develop. The focus of this chapter is on the domain of assessment, primarily as it relates to medical patients.

This chapter begins with a brief discussion of formal aspects of health care settings and introduces the reader to challenges in medical care. Because the medical setting is the context for assessment and the medical patient is the client for assessment, the setting itself must be understood as a critical influence on the practice of psychological assessment. The reader is then introduced to the most typical assessments used in medical settings. The chapter also discusses new directions in assessment. For many patients, medical advances have resulted in the patient's surviving what would previously have been a fatal accident, event, disease, or condition. In this case, the diagnosis becomes one of chronic illness, and the patient's return to work (or normal life) is of interest. Last, because growth of prescription privileges will certainly influence assessment practices in medical settings, the influence of this shift in psychological practice is considered.

The chapter concludes with an overview of ethical and legal issues inherent in the medical setting. Because medical settings have historically operated for the practice of medicine by physicians, there are challenges for psychologists to adhere to their own code of ethics. Ethical imperatives for confidentiality

DOI: 10.1037/14048-017
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

and informed consent, in particular, can pose difficulties.

AN INTRODUCTION TO HEALTH CARE SETTINGS

Medical settings are formalized systems of treatment delivery governed by a variety of federal, state, and local laws; policies; and norms. Psychologists operate in this system of treatment delivery, in that set of laws and policies as well as their own codes of ethics and practice. Thus, medical settings can pose a challenge to psychologists who are more accustomed to defining their own treatment environment.

Three characteristics of medical settings influence health care delivery. First, health care settings were historically organized as practice settings for physicians. Although other health care providers served alongside physicians, the orientation that physicians hold toward patient care dictated the overall goals of diagnosis of symptoms and resolution of medical condition. Although psychologists may be as interested in a patient's emotional state, his or her occupational status, or adjustment to treatment as in symptom resolution, unless these interests dovetail with the overall goal of restoration of health, they are considered secondary.

Second, physicians historically operated as primary decision makers for their patients. Thus, their approach emphasizes control of information gathered, less collaboration with the patient, and immediacy of decision making. Psychologists, in contrast, typically favor a working relationship with clients where information gathered and treatment goals are the product of mutual decision making. A prime example of how this difference can influence assessment in medical settings is that psychological assessment is usually governed by the physician's referral question.

Third, the current climate of cost cutting is also the context of psychologists' work in medical settings. The movement toward integrated health care, with behavioral health a central feature, is clear, but there are significant barriers to complete implementation. One barrier is the physical and organizational structure of medical settings: if psychologists are physically located in an adjacent building rather than on the medical unit, their involvement will be equally peripheral. And if psychologists are not organizationally recognized as equal members of the treatment team, their integration will be hampered. A second barrier is the financial health care environment. Psychologists, as with all health care providers, must demonstrate that their assessments and interventions achieve demonstrable improved patient outcomes and long-term health management in a reduced cost context.

Hospital settings have formal systems of authority and management. The board of directors is ultimately responsible for the hospital, including its financial management. Hospital administrators manage the ongoing activities of hospitals. And medical directors oversee all clinical activities. Professional staff members, including psychologists, report to the medical director. There are typically several organized committees that carry out necessary regulatory tasks: credentials committees review credentials of potential staff, quality assurance committees oversee compliance with standards of practice, and medical records committees maintain an increasingly complex system of patient information and technology.

Hospitals are regulated by a variety of federal and state laws. An example is the Health Insurance Portability and Accountability Act (enacted by Congress in 1996) and the Health Information Technology for Economic and Clinical Health Act (enacted by Congress in 2009 as part of the American Recovery and Reinvestment Act). These federal regulations define a wide array of requirements intended to protect confidential health information while at the same time allowing appropriate exchange of information to benefit patient health outcomes. State laws cover aspects of assessment and practice relevant to licensing and credentialing, thereby establishing the scope of practice and necessary limits to patient confidentiality. Hospital policies often establish procedures by which day-to-day activities are governed.

In hospital settings, psychologists have several roles. They develop and implement treatment plans for patients with primary or comorbid psychosocial problems such as chronic pain or myocardial infarction. They consult with health care providers and systems concerning issues such as staff development, burnout, and worksite health promotion.

In assessment, the focus of this chapter, psychologists formulate assessment plans for individuals, families, health care providers, and the environment. Many psychology associations and publications contain descriptions of hospital-based psychologists and how they have developed their practice, which may be of interest to readers (an example is Roth-Roemer, Kurpius, & Carmin, 1998). A later section in this chapter describes the particular ethical and legal challenges facing psychologists in hospital settings.

APPLICATIONS OF ASSESSMENT IN MEDICAL SETTINGS

Conducting psychological assessments in medical settings requires a multifaceted approach (Allmon et al., 2010). Numerous formal and informal assessment tools have been developed for questions of patient screening (e.g., cognitive functioning), psychopathology, psychosomatic issues, treatment appropriateness, and psychophysiological issues. This section briefly reviews assessment methods in each of these areas. Factors to consider, such as the medical setting as a nontraditional assessment environment, type of assessment tools, and interpretation considerations are discussed. Although this section covers some common assessment tools and methods, it does not provide an exhaustive review of all measures. Interested readers are encouraged to consult recent books on the topic (e.g., Antony & Barlow, 2010; Carlstedt, 2010; Mpofu & Oakland, 2010) and to see especially the related chapters in this volume, such as Chapter 13 regarding psychological assessment in treatment and Chapter 18 concerning outcome assessment in health settings.

Psychometric Issues

Having a thorough understanding of psychometric issues is paramount for any clinical health psychologist. Although understanding reliability and validity issues is essential, clinicians also must understand the distinction between sensitivity and specificity. Sensitivity refers to the extent to which scores from a particular assessment tool correctly detect the presence of a specific condition (e.g., depression) or the proportion of true-positives. Specificity refers to the extent to which test scores correctly identify individuals who do not have a specific condition or the proportion of true-negatives. The relationship between sensitivity and specificity can be graphically depicted by a receiver operating characteristic (ROC) curve; ROC analysis can be used to determine the optimal measure or screening device for particular disorders (Fawcett, 2006).

Because the health care environment generates legal issues, such as disability coverage for chronic pain, psychologists must be particularly careful to select assessment methods that have demonstrated validity and reliability, that address the referral question appropriately, and that can be defended in a legal situation, such as subpoena. Woody (2009) has provided an illustrative overview of the ethical pitfalls that can occur when psychologists are unexpectedly forced to defend their practices in a forensic setting. Although these issues are beyond the scope of this chapter, they must be a central focus for clinical health psychologists. For related discussion of legal issues in testing and assessment, please refer to Chapter 6 in this volume.

Referral Questions

Referral questions direct psychological assessment in the medical setting. Referral sources come from numerous medical disciplines including but not limited to oncology, orthopedics, cardiology, radiology, and primary care physicians. Referral questions involve concerns that may range from treatment adherence issues to screening for depression and anxiety, to substance use issues. Often, referral questions are quite vague, and additional clarification is warranted. As an example, a typical referral question might be worded as "Mrs. Smith is a 43-year-old woman with Type II diabetes. Please evaluate." In these cases, psychologists should attempt to contact the referring provider to determine the exact nature of the question. Returning to the referral, is the issue that Mrs. Smith is noncompliant with her treatment for diabetes, or are there concerns over the effect of her disease on her cognitive abilities?

The type of assessment used will depend on the nature of the referral question. Psychologists often must tailor assessment techniques to accommodate

unique demands of the particular setting because there are numerous factors to consider related to choosing the assessment methods. As an example, assessment in an inpatient setting may be very different than assessment in an outpatient setting. Alternatively, it may not be possible to administer a full neuropsychological battery until a brief cognitive screen is completed.

Some authors have proposed models of assessment in clinical settings that define necessary domains. For example, assessments have been operationalized by biological, affective, cognitive, and behavioral targets (Belar & Deardorff, 2009). In each target, other environmental and sociocultural aspects are taken into consideration. Assessment information in each domain is then integrated into the patient conceptualization.

Understanding cultural considerations is paramount in assessment. For example, research has demonstrated that African Americans are more likely to seek treatment from primary care providers than from mental health providers and have increased rates of exposure to trauma (Snowden & Pingitore, 2002). Along with cultural factors, the type of population using the setting in which the assessment is conducted may be relevant. For example, Roy-Byrne, Russo, Cowley, and Katon (2003) noted that patients receiving care in public sector hospitals experienced higher levels of anxiety and stress, and reported fewer resources and higher levels of unemployment, compared to patients in the private sector.

Clinical Interviews

The clinical interview is the most common method used to gain a comparatively large amount of information in a relatively short amount of time. Coupled with a medical chart review and information from other health care providers, psychologists can assess, form impressions and conceptualizations, and make treatment recommendations based on the information garnered. The clinical interview format is not universal; rather, it must be tailored to the patient population. For example, a clinical interview of a patient in need of a kidney transplant due to endstage renal failure may emphasize compliance issues, including medication management, current or past history of substance use, and availability of caregiver resources and social support. In contrast, clinical interview questions asked of a patient with Type II diabetes may emphasize behavioral indicators including weight management or smoking cessation. The setting, patient population, and potential time available with the patient are factors that influence the content of the clinical interview.

The clinical interview also presents an opportunity to develop rapport and establish a collaborative working relationship with the patient. Psychologists provide empathy, normalize stressors, and build trust while collecting essential information to make appropriate recommendations for treatment. The clinical interview additionally provides opportunity for behavior observation. Behavioral indicators including appearance, demeanor, mood and affect, speech, thought processes, eye contact, and overall relational style can be assessed during the one-to-one interview.

Although it is considered an assessment tool, the clinical interview may serve as a treatment enhancement through the use of *motivational interviewing* (MI; Miller & Rollnick, 2002), an assessment technique that draws from clients' values and motivates clients for change. Randomized controlled trials with health populations have shown MI to be an effective technique for promoting behavioral changes. These changes can be central to the health psychologist working with medical problems of chronic pain, diabetes, hypertension, smoking, and weight management. Thus a clinical interview can focus on patient strengths and emphasize patient motivation to make positive changes that create desired health outcomes.

Mental Status

One of the most common referral questions to psychologists reflects the practitioners' unique role in the health care system as a professional able to knowledgeably evaluate mental status and cognitive impairment. This referral question can be part of many medical issues. Examples include determining whether patients can knowledgeably consent to treatment, whether patients are able to carry out self-care for particular treatment regimens and thus are appropriate candidates for that treatment,

whether patients with chronic medical problems need to be referred to supported living environments, and whether patients are able to continue activities of daily living such as driving a car.

The Mini-Mental State Examination (MMSE; Folstein, Folstein, & Mchugh, 1975) is the most frequently used brief cognitive screening device. The MMSE assesses five areas of cognition: orientation, registration, attention and calculation, recall, and language. Although it is commonly used, Ismail, Rajji, and Shulman (2010) reminded psychologists of several issues that should be taken into consideration with MMSE findings. There may be a ceiling effect with individuals of high premorbid intelligence or education, and it is therefore important to use norms representing the patient population being assessed. Patient age, education, culture, and sensory deficits may produce false-positives, and the MMSE may lack sensitivity to differentiate between mild cognitive impairment (MCI)/dementia and healthy controls (Mitchell, 2009).

The Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) is a brief assessment, administered in fewer than 10 minutes and designed to detect MCI. It assesses cognitive domains of attention and concentration, executive function, memory, language, visuoconstructional skills, conceptual thinking, calculation, and orientation. The MoCA is an improvement over the MMSE in terms of both sensitivity and specificity. The MMSE has a sensitivity of 18%to detect MCI, whereas the MoCA detected 90% of MCI subjects. Similarly, in a group of patients with mild Alzheimer's disease, the MMSE had sensitivity of 78%, whereas the MoCA had 100%.

The Clock Drawing Test (CDT) is another common screening device for cognitive impairment (see Huntzinger, Rosse, Schwartz, Ross, & Deutsch, 1992, for an early version). The CDT is easily administered and takes less time than the MoCA and MMSE. Patients are asked to draw the face of a clock including the numbers and to draw a set of hands set to a particular number. This test taps into cognitive domains—executive functioning, comprehension, planning, visual memory—influenced by dementia. Various research studies have demonstrated interrater reliability, sensitivity, and specificity for predicting consensus diagnosis. The CDT is less culturally biased than other cognitive screening measures and can be used for non-English-speaking populations (Parker & Philp, 2004). However, its wide use has led to subjective and qualitative applications that have been criticized (Nair et al., 2010).

The Saint Louis University Mental Status (SLUMS) examination is a relatively new tool to detect dementia and mild neurocognitive disorder (Tariq, Tumosa, Chibnall, Perry, & Morley, 2006) that assesses orientation, memory, attention, and executive functions. The test authors noted that the SLUMS is an improvement over the MMSE because it taps into more cognitive domains: attention; numeric calculation; immediate and delayed recall; animal naming; digit span; clock drawing; figure recognition, size recognition, size differentiation; and immediate recall of facts from a paragraph. Sensitivity and specificity are similar to that of the MMSE in detecting dementia. However, the SLUMS may be better able to differentiate MNCD than the MMSE. This measure may be helpful in the early detection of cognitive impairments.

Assessment of delirium has received considerably less attention in the research literature but is a common part of mental status assessment. A recent review (Wong, Holroyd-Leduc, Simel, & Straus, 2010) of the ability of various screening methods to diagnose delirium also revealed a range of prevalence of delirium among medical patients. Lower levels (12% to 27%) were found among patients postsurgery; moderate levels (9% to 43%), among geriatric patients; and high levels (63%), among oncology patients. Delirium, compared with dementia, is characterized by rapid onset and alterations in attention and consciousness. The most widely used assessment for delirium is the Confusion Assessment Method (CAM; Inouye et al., 1990). The scale was developed by an expert panel of clinicians who identified nine clinical features of delirium; the resulting diagnostic algorithm is based on the clinical interview and on information obtained during patient observation. Sensitivity was 86% and specificity was 93%.

Several factors should be considered when selecting tools to screen for cognitive issues. Most screening instruments are useful for determining symptom severity and for identifying individuals needing

additional assessment. Scores on screening tools alone cannot be the basis of formal diagnosis. It is also essential to take cultural aspects into account, because numerous tests (e.g., MMSE) are susceptible to cultural influences from variables such as educational level and ethnicity (Parker & Philp, 2004). Unfortunately, because these tools are short and time sensitive, there is opportunity for misuse. Additionally, as noted later in the discussion pertaining to ethical considerations, sometimes psychologists simply do not have appropriate measures. As an example, if a psychologist in an urban hospital were asked to assess a refugee from Botswana for cognitive impairment, there are few, if any, measures that would adequately take this person's cultural context into account.

Psychopathology and Adjustment

The most common disorders evaluated in medical settings are related to symptoms of depression, anxiety, and substance abuse (Spitzer, Kroenke, Williams, & Löwe, 2006; Spitzer, Kroenke, Williams, & the Patient Health Questionnaire Primary Care Study Group, 1999). Major depression and symptoms of depression are common among patients presenting in medical settings. Pearson et al. (1999) demonstrated that 20% of patients identified as high utilizers of medical care experienced major depression or major depression in partial remission. Among primary care settings, the prevalence of some form of depression has been reported to range between 15% and 22% (Jarrett, 2009).

The Beck Depression Inventory—II (BDI–II; Beck, Steer, & Brown, 1996) is a widely used screening instrument that identifies the severity of depressive symptoms. Participants respond to groups of statements that describe symptoms such as feelings of worthlessness and disturbances in eating and sleeping. The BDI–II has acceptable psychometric qualities among medical populations and minority populations (Grothe et al., 2005). A short BDI–II with sensitivity of 100% and specificity of 83% is also available (Furlanetto, Mendlowicz, & Bueno, 2005). The Center for Epidemiologic Studies—Depression scale (CES–D; Radloff, 1977) is an alternative measure of depressive symptoms among chronically ill samples. The rationale behind the CES–D is that traditional depression inventories rely heavily on somatic indices of depression and thus will be a biased measure among patients who are physically ill. Sensitivity for the CES–D ranges from 75% to 93% and specificity from 73% to 87% (Watson & Pignone, 2003).

Along with depressive disorders, suicidal ideation and self-injurious behaviors are important targets for assessment because knowing if patients are at increased risk for suicide is critical. Risk factors include one or more previous suicide attempts; history of psychiatric illness (e.g., major depressive disorder); alcohol dependence; family history of suicide; and stressful life events (Lake & Baumer, 2010). Screening for suicidal intent can be completed with assessment tools already described or by asking specific questions: "Are you having thoughts of death? Are you wishing that you were dead?" Using direct questions has been found to demonstrate sensitivity and specificity for identifying individuals contemplating suicide (Gaynes et al., 2004).

Anxiety disorders are highly prevalent in medical settings. In a study of 965 primary care patients, approximately 20% had at least one anxiety disorder followed by posttraumatic stress disorder (PTSD; 8.6%), generalized anxiety disorder (GAD, 7.6%), and panic disorder (6.8%) (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007). Several assessment methods detect symptoms of anxiety. The seven-item Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006) has sensitivity (89%) and specificity (82%). The Patient Health Questionnaire was developed from the Primary Care Evaluation of Mental Disorders (Spitzer et al., 1999). It consists of items that assess anxiety, depression, alcohol use, somatoform, and eating disorders and has demonstrated good validity with specificity of 88% for major depression.

Anxiety disorders and depression often present along with somatic complaints (Kroenke et al., 2007). Therefore, it is important to take comorbid medical conditions into account in assessment. Maizels, Smitherman, and Penzien (2006) suggest that several screening devices be used when psychologists screen for anxiety and depression because some measures do not adequately capture cognitive symptoms and focus more on somatic symptoms.

Individuals presenting in medical settings often have trauma histories resulting in symptoms of PTSD. In medical settings, the prevalence for individuals meeting full or partial criteria for PTSD ranges from 9% to 25% (Gillock, Zayfert, Hegel, & Ferguson, 2005). A frequently used assessment of posttraumatic symptoms is the Posttraumatic Stress Disorders Checklist (PCL; Blanchard, Jones-Alexander, Buckley, & Forneris, 1996). The PCL assesses three symptom clusters of PTSD: re-experiencing, avoidance, and increased arousal. There are formats for persons who have completed military service, for those responding to general stress, and those who experienced a specific trauma. The PCL can be scored with an overall symptom severity index, a cut score, or levels of symptoms based on clusters. Research indicates that the PCL has good sensitivity (78% to 94%) and specificity (68% to 71%) in primary care settings.

Traditionally, psychologists have not worked in emergency room settings although their role is increasing in these environments that present unique challenges for assessment. For example, symptoms of panic disorder (PD; e.g., shortness of breath, chest pain) are similar to those of myocardial infarction or coronary artery disease (CAD). Lynch and Galbraith (2003) demonstrated that PD goes undiagnosed in emergency rooms over 95% of the time. Possible reasons for the inaccurate diagnosis may be linked to patients' beliefs about suffering from a major medical illness, overlap of symptoms between PD and medical illness, and physicians' motivation to determine a medical cause of presenting symptoms. The most common medical differential between CAD and PD is coronary angiography, a costly and invasive procedure. Expanded assessment in emergency rooms could reduce misdiagnoses, reduce the unnecessary use of procedures, and ultimately reduce health care costs.

Substance use and abuse is associated with numerous health problems. The Alcohol Use Disorders Identification Test (AUDIT) is a measure developed by the World Health Organization (WHO) to assess problematic or risky alcohol consumption (Saunders, Aasland, Babor, de la Fuente, & Grant, 1993). The test's authors noted that the AUDIT can be used with medical patients, including individuals with medical disorders such as pancreatitis, cirrhosis, gastritis, tuberculosis, neurological disorders, and cardiomyopathy. The questionnaire measures recent alcohol use, alcohol dependence symptoms, and alcohol-related problems. A review of the literature indicates that the AUDIT has adequate levels of sensitivity and specificity in various settings including emergency rooms, inner-city medical clinics, family practice clinics, and hospital inpatient units (Allen, Litten, Fertig, & Babor, 1997).

An older alcohol screening device is the CAGE questionnaire (Ewing, 1984; Mayfield, McLeod, & Hall, 1974). The CAGE mnemonic consists of four questions related to whether one feels the need to *cut* down on drinking, whether other people are *annoyed* with the individual's drinking, whether one feels *guilty* about drinking, and whether that individual drinks to offset a hangover in the morning (i.e., has an *eye-opener*). One study suggests that the CAGE with past-year wording, compared with the original wording, was less sensitive (57% vs. 77%) yet more specific (8% vs. 59%; Bradley, Kivlahan, Bush, McDonell, & Fihn, 2001). Related assessments are for drug use and tobacco or nicotine use. Duration and frequency of use, drug dependence, and motivation to quit are also relevant assessment targets.

Pain

Psychologists who work in medical settings will inevitably encounter patients presenting with reported pain even if pain is secondary to the medical diagnosis. Although patients experience numerous types of pain (e.g., neuropathic, nociceptive), the broad categories of *acute* and *chronic* pain are most commonly used. Chronic pain is pain lasting more than 6 months that causes interference in daily life activities. Acute pain is pain that has a sudden onset and fewer than 6 months duration.

The prevalence of pain symptoms among medical populations is high. For example, approximately 50% of breast cancer patients report significant pain levels (Tasmuth et al., 1995). It is important for clinical health psychologists to have an understanding of pain symptoms among numerous disorders, including rheumatologic disorders, various cancers, chronic fatigue syndrome, gastrointestinal disorders,

neurological conditions, and lupus (Boothby, Kuhajda, & Thorn, 2003). For the purpose of this chapter, chronic pain as a particular presentation is emphasized later; however, pain measurement is similar for acute versus chronic pain and for pain as a symptom among other issues to be assessed.

Pain experience is multidimensional; pain domains include intensity and duration; cognitions (e.g., beliefs and perceptions); interference with daily living activities (e.g., walking, sitting in the car); and coping behaviors (see Turk & Melzack, 2001). The most common early assessment tools for pain were the visual analogue scale (VAS), numerical rating scales, and verbal rating scales. The VAS is a subjective measure of pain perceptions. Patients mark a line along a continuum, usually a horizontal line with two anchor words on each side that represent pain at opposite dimensions. Common anchors are "pain that is barely noticeable" and "the worst pain I have ever experienced."

The McGill Pain Questionnaire (MPQ; Melzack, 1975, 1987) is a frequently used multidimensional assessment. The MPQ consists of 20 groups of adjectives that represent current pain experience (the short form has 15) on sensory, affective, and evaluative dimensions. The MPQ can be scored for a pain rating index, the number of words chosen, and the present pain intensity. The popularity of the MPQ may be due to its multidimensional assessment. Piotrowski (2007) reviewed the psychological literature for pain assessments and found the MPQ to be the most common measure, followed by the Multidimensional Pain Inventory (MPI; Kerns, Turk, & Rudy, 1985). The latter measure, the MPI, considers the influence of pain on individuals' activities of daily living and the responses of significant others such as family members to the display of pain behaviors (e.g., moaning, sighing, holding body parts).

One challenge is assessing pain among individuals with cognitive impairment. Verbal rating scales may cause confusion of pain with other constructs, such as depression (Stolee et al., 2005). Behavioral measures sensitive to affective symptoms and physiological measures may be the best choice for individuals with cognitive impairment or communication issues. When working with members of this population, it is important to use several assessment modalities including self-report measures, collateral (e.g., family) information, assessment of functional impairment, and other psychological measures (e.g., depression, anxiety).

In the age of electronics and various handheld devices, technological advances will certainly be integrated into the assessment of psychological phenomena and medical issues. One study found that patients preferred tracking mood, pain, activity interference, and medications electronically compared to paper-and-pencil methods (Marceau, Link, Jamison, & Carolan, 2007). However, additional research is needed to determine validity of electronic assessment methods especially in terms of their comparability with traditional methods. Also, devices have unique problems, such as losing power, transmission difficulties, and patients' not following through with completing assignments, all of which pose challenges to accurate assessment.

Treatment Appropriateness

Psychological assessment provides a critical contribution to medical decision making concerning the acceptability of patients for surgical procedures such as organ transplantation. The most common transplants are for organs (e.g., heart, lung, liver, kidney, and pancreas) and individuals requiring a stem cell transplant (more typically referred to as bone marrow transplantation). Psychosocial evaluations for transplantation have two primary foci: identifying the patient's level of understanding and evaluating factors that influence pre- or postoperative outcomes (e.g., Allmon et al., 2010).

Olbrisch, Benedict, Ashe, and Levenson (2002) identified several dimensions of a psychosocial assessment for transplant consideration. First, the patient's medical history (e.g., previous diagnoses or hospitalizations) and history of psychopathology must be established. Additionally, the patient's level of understanding of his or her condition and treatment, motivation for transplant, and outcome expectations are important targets. Psychologists also assess for risk factors related to noncompliance and poor transplant outcomes, including substance abuse issues, severe psychopathology, and history of treatment noncompliance.

A key component of the transplant assessment is the level of support that transplant candidates expect on returning home. Because of the invasiveness of the procedure and the long recovery time after transplantation, it is critical that patients have a primary caregiver and supportive others in their home environment. Therefore, if possible, caregivers' mental and physical status should be assessed to determine their suitability. Other aspects of the psychosocial assessment include a thorough clinical interview, review of medical records, mental screen, personality profiles, and coping styles.

Collins and Labott (2007) identified measures specific to the transplant evaluation, among them the Psychosocial Assessment of Candidates for Transplantation (Olbrisch, Levenson, & Hamer, 1989) and the Transplant Evaluation Rating Scale (Twillman, Manetto, Wellisch, & Wolcott, 1993). Comprehensive tools to assist clinicians through the transplant process also have been developed. The Structured Interview for Renal Transplantation (Mori, Gallagher, & Milne, 2000) has sections to guide psychologists through the transplant assessment: background/demographics, understanding of illness, education/economic status, brief family history, coping/personality style, psychiatric history, and mental status. The Millon Behavioral Medicine Diagnostic (Millon, Antoni, Millon, Minor, & Grossman, 2006) was designed to assess psychological factors that influence treatment issues and is frequently used for pretransplant evaluations. The inventory assesses domains including response patterns, negative health habits, psychiatric indicators, coping styles, stress moderators, and treatment prognostics.

Psychophysiological Assessment

Psychophysiological methods assess patients' physiological reactions to stressors and other conditions. With these assessments, patients can also be taught to regulate bodily responses in the autonomic nervous system. Psychophysiological assessments have been used to evaluate (and treat) a number of medical issues including chronic pain, cancer pain, phonatory disorders, bowel and bladder disorders, migraine and tension headaches, and chronic obstructive pulmonary disease (COPD).

Electromyography, a common assessment method, indirectly measures muscle contraction through electrical activity in muscles. Electrodes are placed at various points on the body to gauge muscle activity. Assessing skin temperature is another common method. When blood vessels dilate, additional blood flows through the body, thus warming the tissues around blood vessels including the skin. Sedlacek and Taub (1996) noted that 80% to 90% of patients with Raynaud's disease, a disease characterized by vasospastic attacks, can use techniques such as monitoring skin temperature for successful treatment outcomes. Along with skin temperature, skin conductance can be used to measure sweat gland activity among medical patients. Skin conductance, also referred to as *galvanic skin response*, utilizes small electrodes placed on the skin and measures the amount of electrical activity passing through the skin (Peek, 2003). Actigraphy is a noninvasive assessment method for measuring motor activity in individuals experiencing numerous medical issues including sleep disturbance, delirium, and dementia (Ancoli-Israel et al., 2003). This instrument is usually worn on a patient's wrist and assesses motor functioning on an ongoing basis.

A promising new advancement in psychophysiological assessment targets heart rate variability. Heart rate variability occurs when the interval between heartbeats fluctuates, a situation linked to medical and psychological conditions including cardiovascular disease, metabolic syndrome, depression, and anxiety (McGrady, 2007). Patients are taught breathing exercises to regulate sympathetic responses and to decrease heart rate variability with ongoing psychophysiological assessment. A recent review of the literature suggests that heart rate variability biofeedback techniques may also be useful for asthma, cardiovascular disease, COPD, heart failure, fibromyalgia, and PTSD (Wheat & Larkin, 2010).

With increasing technological advances, clinicians in medical centers can work with patients' assessment results and treat individuals over the internet (Olsson, El Alaoui, Carlberg, Carlbring & Ghaderi, 2010). Using the Internet as a therapist–patient meeting ground reduces costs of hospital visits and allows more access for patients who travel

long distances for hospital visits. Although this area of research and application is burgeoning, the efficacy of the intersection of psychophysiological assessment techniques and Internet-based treatment must be established as an effective treatment modality.

NEW DIRECTIONS

Screening

Psychological screening, compared with a full diagnostic battery, is much briefer and usually conducted in the timeframe of the initial appointment. Screening assessments do not allow the specificity of a longer battery of tests but may provide preliminary evidence for the presence of a disorder. Kessler (2009) identified several criteria to consider when screening instruments are used or developing screening programs:

- Which patient population is to be screened? Examples are persons admitted from the emergency room, children under age 5, and adults over 65.
- What clinical domains are to be covered, and what measures will be used? Examples are depression, pain, and cognitive impairment.
- How much time is available to collect the data? A patient in a busy clinic may have only a few minutes between appointments, whereas a full day spent at a teaching hospital might give a patient an hour or more between appointments.
- How is the procedure going to happen in the setting, who does it, and where? The realities of crowded medical settings are such that screening may happen at a patient's bedside in competition with television and visitors, in a hallway where staff are walking by, or in a waiting room.
- What format is the most efficient and most valid and reliable? Time is at a premium in medical settings and brevity is paramount; therefore, the use of quick but psychometrically sound instruments is critical. Psychologists can choose from traditional paper-and-pencil assessment, face-to-face assessment, behavioral observation, or electronic formats to gather information.

More health care systems are moving toward an electronic kiosk system to gather information. Electronic tablets or notebooks are assessment methods that are used while patients are in waiting rooms, and many psychological screening measures can be integrated within medical information gathering. Several researchers have noted the benefits of using electronic methods for assessment (Kessler, 2009; Provenzano, Fanciullo, Jamison, McHugo, & Baird, 2007). Electronic assessment saves time for staff members, scoring is performed automatically, and information can be transferred to a central database where patient information can be integrated as well as made available to other providers. Visual depictions of symptoms by means of graphs or charts provide an interpretative dimension and can be helpful when assessment results are given to patients. Statistical analysis can yield information on meaningful differences in patients across time. In light of the ease of implementing technological advances, including electronic assessments, it is possible that psychometric considerations are overlooked or minimized; electronic assessments must be used only for their designed intention (Caraceni, Brunelli, Martini, Zecca, & De Conno, 2005).

In the technological future, psychological assessment will increasingly rely on telehealth methods, as is currently done in the Veterans Affairs system. This type of assessment presents unique challenges for clinicians. The briefer the measure, the more important it is to spend time with the patient for follow up. Telehealth, as with other Internet-based methods of assessment and treatment, must take into account issues related to diversity, including ethnic and racial differences, and consider measures sensitive to social class (socioeconomic status, education, income). Additionally, all the limitations of the internet apply to web-based assessment: the possibility of misunderstanding due to the missing nonverbal aspect of communication, the technological concerns of lost data and terminated connections, and the potential for reduced confidentiality.

Kessler (2009) identified and summarized additional ethical considerations when screening devices are used. Perhaps the most important of these considerations is that resources must be available to detect and interpret "positive" screens and a treatment plan

must be ready for implementation. Screening unlinked with treatment referral can result in a failure to meet client needs and the likelihood of misusing data. Additionally, there are potential negative effects of screening measures: limits to confidentiality, the standard for a positive screen (false positives vs. false negatives), and the need to resist establishing a diagnosis from a single screening measure.

Prescription Privileges

Psychologists' right to prescribe medication continues to be a contested issue. Currently, only two states—New Mexico and Louisiana—allow practicing psychologists to prescribe medications. There are arguments for and against prescriptive authority for psychologists (for a review of the issues, see McGrath, 2010), and future debate will be likely as psychology expands its presence in integrated care models. For psychologists with the authority to prescribe medication, there are numerous areas in psychological assessment that must be considered.

Whether a psychologist has prescription privileges will greatly influence assessments conducted in medical settings. Ally (2010) noted several areas that psychologists need to consider when they perform assessments focused on both psychological and medication treatment evaluations. An in-depth assessment of medical history including vital signs, allergic reactions to medications, past or current medication usage, previous medical evaluation, recent laboratory work, and issues related to pregnancy (including breastfeeding) must be performed. For these assessments, psychologists need equipment including pen light, reflex hammer, stethoscope, sphygmomanometer, thermometer, and scale. Psychologists would also need the appropriate training to use this equipment and interpret the findings. Gruber (2010) discussed psychologists having knowledge of medical terminology including acronyms for drug names; information on dosage and routes of administration; and supplemental knowledge concerning over-the-counter medication, vitamins, supplements, and homeopathic preparations.

LeVine and Foster (2010) outlined a set of domains additional to assessment when prescription is considered. Assessing the therapeutic potential of medications is a critical component. Other topics for continued assessment include weight gain or loss, lipid and insulin levels, blood pressure, sexual performance, irritability, and mental health issues (e.g., anxiety). Furthermore, psychologists should be knowledgeable of possible medication contraindications before prescribing medications. For example, substance abuse issues, medication noncompliance, high blood pressure, or other conditions that have a deleterious effect on the patient should be assessed. At-risk behaviors including drug dependence potential and drug seeking behaviors also should be considered.

Whether psychologists have prescribing privileges or not, practitioners working in medical settings must have an understanding of medications. With that said, psychologists practice within an ethical framework when they restrict their work to those activities that fall within the limits of their own competency regarding recommendations for medication issues. A potential challenge for prescribing psychologists is taking assessment information into account, forming a case disposition, and determining the appropriate treatment modality including psychotherapy, medication, or both medication and psychotherapy (Ally, 2010; LeVine & Wiggins, 2010).

As drug companies recognize the increasing influence of psychologists in prescription decision making, psychologists will face the same types of ethically challenging pressures historically confronted by physicians. National medical associations have taken stricter positions recently against physicians accepting gifts or other benefits from drug companies, but psychological associations have not yet confronted this issue. As an example, a pediatric psychologist may encourage his patients' parents to ask their physicians for a particular medication for their child who has been diagnosed with an attention deficit disorder. Although the psychologist may well believe that a particular medication has performed well for clients under his care, his scope of practice, legal standing, and ethical competence can be easily threatened.

Vocational and Career Assessment

Illnesses once considered terminal are now often viewed as chronic in nature. Cancer is an excellent

example: With early detection and appropriate treatment, individuals now diagnosed with cancer are many more times likely to survive and return to "normal living" than even a decade ago. Consequently, many individuals are living with a chronic health condition but continue to experience functional deficits that influence their ability to perform daily activities, including employment. For example, individuals with HIV/AIDS face a number of impairments in daily life that can influence job-related activities (Anandan, Braverman, Kielhofner, & Forsyth, 2006). Additional medical conditions that impede vocational functioning are chronic pain, epilepsy, traumatic brain injury, and multiple sclerosis.

After experiencing a debilitating medical condition, patients are often not able to return to previous employment or careers. Therefore, a new area of assessment for health psychologists is considering how patients' current vocational interests intersect the range of possible employment options. By assessing vocational interests of patients managing chronic health conditions, psychologists can direct patients to realistic and satisfying career options. The Strong Interest Inventory (SII; Donnay, Morris, Schaubhut, & Thompson, 2004), a popular tool in career interest assessment, is designed to assess basic interests, occupational themes, and personal styles. There is also an administrative index that is useful for determining problematic profiles or random responding. (See several chapters in the counseling psychology section of this volume of the handbook, especially Chapter 19 on the assessment of interests, where additional information on the Strong and the Self-Directed Search [SDS; Holland, 1994] may be found.)

The SDS is a self-administered inventory that examines interests based on Holland's (1994) theory of vocational personality types. Holland types—Realistic, Investigative, Artistic, Social, Enterprising, and Conventional—are based on the premise that vocational interests stem from personality style, and the most satisfying occupations are those that match an individual's style. For example, a Realistic type enjoys working with things (e.g., tools, machines) more than with people or working outdoors. Perhaps a Realistic person who previously handled construction machinery can no longer do so because of medication that impairs balance. However, directing this person to a Social occupation—working with people to be helpful around educational and social issues—may not result in employment. An alternate and better strategy is to consider a variety of Realistic occupations that would allow the patient to reengage in work in the preferred area. Both the SII and SDS can be effective tools in assisting medical patients in changing their careers due to chronic health concerns.

Posttraumatic Growth

Assessment techniques in medical settings have traditionally focused on psychopathology or negative sequelae associated with medical disorders. Assessing dysfunction, rather than functionality or positive gains, has been a hallmark feature of the medical model. However, a considerable research base has developed that documents the repeated finding that very difficult, and even traumatic, life events may prompt significant personal growth among individuals who experience them. Posttraumatic growth has been defined as experiencing positive psychological gains that result from traumatic or stressful life experiences (Tedeschi & Calhoun, 1995). Individuals experiencing traumatic medical conditions including cancer, spinal cord injury, bone marrow transplant, multiple sclerosis, and burns have reported themes of growth (for a review, see Barskova & Oesterreich, 2009).

The Posttraumatic Growth Inventory (PTGI; Tedeschi & Calhoun, 1996) is a widely used measure to assess positive psychological changes, including medically ill patients. The PTGI defines growth in five domains: interpersonal relationships, life philosophy/perspective, new life directions, spiritual outlook, and personal strength. Each of these domains represents areas in which individuals have reported growth. The five-factor structure of the PTGI was supported with confirmatory factor analysis (Brunet, McDonough, Hadd, Crocker, & Sabiston, 2010). Another measure, the Benefit Finding Scale (Tomich & Helgeson, 2004), includes domains of personal priorities, daily activities, family, worldviews, relationships, career, religion, and social contact. Along with standardized assessment instruments, open-ended questions can obtain

patients' self-reported growth. An advantage of an open-ended format is identification of domains that are not typically measured by traditional growth assessments. Thus, growth related to health or medical benefits (e.g., learning more about one's illness, learning how to interact with doctors, improved health) is often not assessed by standardized measures (Tallman, Shaw, Schultz, & Altmaier, 2010).

ETHICAL CHALLENGES

Returning to the theme of the first section, working in health care settings is an exciting opportunity for psychologists. Nationally, the movement toward integrated primary care is one in which psychologists can easily engage. Personally, working with issues of health, illness, death, and dying are challenging yet satisfying ways to contribute to client welfare specifically and social welfare more generally. However, health care settings embody unique challenges as far as the ethical principles and standards that psychologists use to guide their practice (American Psychological Association [APA], 2010).

Competence

Competence (Ethical Standard 2.01; APA, 2010) restricts psychologists' practice to areas in which they have been trained and supervised unless particular emergency situations apply. An additional boundary on psychologists' actions in medical setting is the legal liability resulting from practicing outside one's defined scope of practice, failure to consult and refer, and malpractice. Psychologists must continually strive to maintain and expand the boundaries of their competence through appropriate training and supervision. Additionally, psychologists who work with persons of different backgrounds (e.g., race, social class, religion, disability) require additional competence (Ethical Standard 2.10b).

Assessment

The particular topic of this chapter, assessment, is the focus of an entire ethical standard (Ethical Standard 9; APA, 2010), reminding the profession that tasks of assessment (selecting assessment method, obtaining informed consent, interpretation, and recommendations) are rife with ethical challenges. As noted in the section on screening, psychologists use assessment methods for which there is established test score validity and reliability with the population being tested (Ethical Standard 9.02b; APA, 2010). As attractive as computerized screening methods are to physicians and other medical personnel, psychologists cannot condone these applications without attention to psychometric issues.

Assessment in medical settings is always completed in response to a referral question or to a regular practice of screening, which leads to consideration of the results of the assessment. Is the psychologist responsible only to the referring physician? Does the psychologist need to ensure that the patient also receives an interpretation of the results? Belar and Deardorff (2009) summarized the ethical threats in routine assessment scenarios. The health psychologist must select proper comparison data for patients who are medically ill as opposed to psychiatric populations (Ethical Standards 9.02a and 9.02b; APA, 2010), must consider the risk of inappropriate use of the test results by nonpsychologist health care providers (Ethical Standards 9.06 and 9.07; APA, 2010), and must consider how the patient is to receive an interpretation of the test results (Ethical Standard 9.10; APA, 2010).

Confidentiality

The multidisciplinary medical setting also challenges ethical standards relating to confidentiality. Psychologists who are accustomed to gathering and maintaining clinical information in their own files are surprised by the amount of information available on patients through their electronic records to anyone in the setting with a legitimate need to access the information. Health care settings also can challenge the confidentiality of psychological information when medical records are disclosed to outside providers.

Psychologists must strive to meet standards of confidentiality even when these standards are pressured by the setting. For example, if a patient is to be tested by a psychologist and the information shared with the patient's treatment team, then the patient should be informed of what material will be given to the team. If a patient is not in a private location, the psychologist should attempt to relocate the

testing session. Finally, when limits on confidentiality are of concern to the psychologist, he or she can discuss the influence of these limits with the patient and document patient concerns if necessary.

Informed Consent

Frequently, patients arrive for testing with little knowledge of the purpose of the assessment. It is not uncommon for a patient to present himself or herself at an appointment with no other information than "my doctor told me to come." Thus, psychologists must attend to ethical standards of informed consent at the beginning of the appointment. Ethical Standard 9.03 (Informed Consent in Assessments; APA, 2010) requires that psychologists provide information sufficient to enable patients to understand the nature and purpose of the assessment, and the limits to confidentiality of the results.

Informed consent, as a term, contains the two necessary components of this action. The first is that the patient is fully *informed*, using language that he or she can understand, about the assessment to which the patient is consenting. Many hospitals use consent forms that have such an advanced reading level that some patients may give consent without understanding their decision. Alternatively, in response to legal concerns about abbreviated information, settings may provide such lengthy and wordy documents that the patient feels overwhelmed. In either case, the psychologist must review the information with the patient and document that the patient understands. The second important concern is *consent*, which indicates that the patient freely and voluntarily agrees to the assessment. Recent shifts in health care that empower the patient to act as a consumer and exercise autonomy have not really eliminated the likelihood that patients agree to whatever their physician or health care provider suggests.

The psychologist can assess the patient's truly informed consent, obtaining the patient's perspective on what content the tests will cover, why the tests were ordered, and the uncertainty caused by the consent process itself as well as the assessment. In assessment for a transplant, for example, one possible outcome is that the patient is not approved for the treatment. Thus, the patient understands that the psychologist has input in a medical decision but may need to discuss misconceptions about what factors will be considered in the assessment.

CONCLUSION

The stakes in health care are high. Political agendas for reform and cost containment, the rising price of medication and equipment, increased expectations from consumers, and even the "graying" of the American population have resulted in radical changes in health care. As Kirchner et al. (2010) noted, innovation in medical settings, particularly in primary care, involves many change agents and stakeholders. More than ever, assessment in medical settings is a stimulating area in which to be involved and provides psychologists with significant access into improved medical care. This chapter has considered the health care context to assessment, outlined major targets of assessment along with frequently used measures, considered likely new directions in assessment, and outlined ethical challenges that have been and will be confronted by psychologists. Our hope as authors is that readers have been energized to enter this practice setting.

References

Allen, J. P., Litten, R. Z., Fertig, J. B., & Babor, T. (1997). A review of research on the Alcohol Use Disorders Identification Test (AUDIT). *Alcoholism: Clinical and Experimental Research, 21*, 613–619. doi:10.1111/j.1530-0277.1997.tb03811.x

Allmon, A., Shaw, K., Martens, J., Yamada, T., Lohnberg, J., & Schultz, J. . . . Altmaier, E. M. (2010, Spring). Organ transplantation: Issues in assessment and treatment. *The Register Report*, *36*, 10–17. Retrieved from http://www.nationalregister.org/benefits_publications.html

Ally, G. A. (2010). Nuts and bolts of prescriptive practice. In R. E. McGrath & B. A. Moore (Eds.), *Pharmacotherapy for psychologists: Prescribing and collaborative roles* (pp. 71–87). Washington, DC: American Psychological Association. doi:10.1037/12167-004

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

Anandan, N., Braverman, B., Kielhofner, G., & Forsyth, K. (2006). Impairments and perceived competence

in persons living with HIV/AIDS. *Work: A Journal of Prevention, Assessment, and Rehabilitation*, *27*, 255–266. Retrieved from http://iospress.metapress.com/content/103190

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., & Pollak, C. P. (2003). The role of actigraphy in the study of sleep and circadian rhythms. *Sleep: Journal of Sleep and Sleep Disorders Research*, *26*, 342–392.

Antony, M. M., & Barlow, D. H. (Eds.). (2010). *Handbook of assessment and treatment planning for psychological disorders* (2nd ed.). New York, NY: Guilford Press.

Barskova, T., & Oesterreich, R. (2009). Post-traumatic growth in people living with a serious medical condition and its relations to physical and mental health: A systematic review. *Disability and Rehabilitation, 31*, 1709–1733. doi:10.1080/09638280902738441

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Belar, C. D., & Deardorff, W. W. (2009). *Clinical health psychology in medical settings: A practitioner's guidebook* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/11852-000

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist (PCL). *Behaviour Research and Therapy, 34*, 669–673. doi:10.1016/0005-7967(96)00033-2

Boothby, J. L., Kuhajda, M. C., & Thorn, B. E. (2003). Diagnostic and treatment considerations in chronic pain. In L. M. Cohen, D. E. McChargue, & F. L. Collins Jr. (Eds.), *The health psychology handbook: Practical issues for the behavioral medicine specialist* (pp. 229–251). Thousand Oaks, CA: Sage.

Bradley, K. A., Kivlahan, D. R., Bush, K. R., McDonell, M. B., & Fihn, S. D. (2001). Variations on the CAGE alcohol screening questionnaire: Strengths and limitations in VA general medical patients. *Alcoholism: Clinical and Experimental Research, 25*, 1472–1478. doi:10.1111/j.1530-0277.2001.tb02149.x

Brunet, J., McDonough, M. H., Hadd, V., Crocker, P. R. E., & Sabiston, C. M. (2010). The Posttraumatic Growth Inventory: An examination of the factor structure and invariance among breast cancer survivors. *Psycho-Oncology, 19*, 830–838. doi:10.1002/pon.1640

Caraceni, A., Brunelli, C., Martini, C., Zecca, E., & De Conno, F. (2005). Cancer pain assessment in clinical trials. A review of the literature (1999–2002). *Journal of Pain and Symptom Management, 29*, 507–519. doi:10.1016/j.jpainsymman.2004.08.014

Carlstedt, R. A. (Ed.). (2010). *Handbook of integrative clinical psychology, psychiatry, and behavioral medicine: Perspectives, practices, and research*. New York, NY: Springer.

Collins, C. A., & Labott, S. M. (2007). Psychological assessment of candidates for solid organ transplantation. *Professional Psychology: Research and Practice, 38*, 150–157. doi:10.1037/0735-7028.38.2.150

Donnay, D. A. C., Morris, M. L., Schaubhut, N. A., & Thompson, R. C. (2004). *Strong Interest Inventory manual: Research, development, and strategies for interpretation*. Mountain View, CA: Consulting Psychology Press.

Ewing, J. A. (1984). Detecting alcoholism: The CAGE questionnaire. *JAMA*, *252*, 1905–1907.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874. doi:10.1016/j.patrec.2005.10.010

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189–198. doi:10.1016/0022-3956(75)90026-6

Furlanetto, L. M., Mendlowicz, M. V., & Bueno, J. R. (2005). The validity of the Beck Depression Inventory-Short Form as a screening and diagnostic instrument for moderate and severe depression in medical inpatients. *Journal of Affective Disorders, 86*, 87–91. doi:10.1016/j.jad.2004.12.011

Gaynes, B. N., West, S. L., Ford, C. A., Frame, P., Klein, J., & Lohr, K. N. (2004). Screening for suicide risk in adults: A summary of the evidence for the U.S. Preventative Services Task Force. *Annals of Internal Medicine*, *140*, 822–835.

Gillock, K. L., Zayfert, C., Hegel, M. T., & Ferguson, R. J. (2005). Posttraumatic stress disorder in primary care: Prevalence and relationships with physical symptoms and medical utilization. *General Hospital Psychiatry, 27*, 392–399. doi:10.1016/j.genhosppsych.2005.06.004

Gray, G. V., Brody, D. S., & Johnson, D. (2005). The evolution of behavioral primary care. *Professional Psychology: Research and Practice, 36*, 123–129. doi:10.1037/0735-7028.36.2.123

Grothe, K. B., Dutton, G. R., Jones, G. N., Bodenlos, J., Ancona, M., & Brantley, P. J. (2005). Validation of the Beck Depression Inventory—II in a low-income African American sample of medical outpatients. *Psychological Assessment, 17*, 110–114. doi:10.1037/1040-3590.17.1.110

Gruber, A. R. (2010). Psychologists in primary care. In R. E. McGrath & B. A. Moore (Eds.), *Pharmacotherapy for psychologists: Prescribing and collaborative roles* (pp. 137–187). Washington, DC: American Psychological Association. doi:10.1037/12167-009

Holland, J. L. (1994). *The Self-Directed Search: Professional manual—Form R*. Odessa, FL: Psychological Assessment Resources.

Huntzinger, J. A., Rosse, R. B., Schwartz, B. L., Ross, L. A., & Deutsch, S. I. (1992). Clock drawing in the screening assessment of cognitive impairment in an ambulatory care setting: A preliminary report. *General Hospital Psychiatry, 14*, 142–144. doi:10.1016/0163-8343(92)90040-H

Inouye, S. K., van Dyck, C. H., Alessi, C. A., Balkin, S., Siegal, A. P., & Horwitz, R. I. (1990). Clarifying confusion: The Confusion Assessment Method. A new method for detection of delirium. *Annals of Internal Medicine, 113*, 941–948.

Ismail, Z., Rajji, T. K., & Shulman, K. I. (2010). Brief cognitive screening instruments: An update. *International Journal of Geriatric Psychiatry, 25*, 111–120. doi:10.1002/gps.2306

James, L. C., & Folen, R. A. (Eds.). (2005).*The primary care consultant: The next frontier for psychologists in hospitals and clinics*.Washington, DC: American Psychological Association. doi:10.1037/10962-000

Jarrett, E. M. (2009). The primary care consultant toolkit: Tools for behavioral medicine. In L. C. James & W. T. O'Donohue (Eds.), *The primary care toolkit: Practical resources for the integrated behavioral care provider* (pp. 133–167). New York, NY: Springer. doi:10.1007/978-0-387-78971-2_11

Kerns, R. D., Turk, D. C., & Rudy, T. E. (1985). The West Haven–Yale Multidimensional Pain Inventory (WHYMPI). *Pain, 23*, 345–356. doi:10.1016/0304-3959(85)90004-1

Kessler, R. (2009). Identifying and screening for psychological and comorbid medical and psychological disorders in medical settings. *Journal of Clinical Psychology, 65*, 253–267. doi:10.1002/jclp.20546

Kirchner, J., Edlund, C. N., Henderson, K., Daily, L., Parker, L. E., & Fortney, J. C. (2010). Using a multilevel approach to implement a primary care mental health (PCMH) program. *Families, Systems, and Health, 28*, 161–174. doi:10.1037/a0020250

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*, *146*, 317–325.

Lake, C. R., & Baumer, J. (2010). Academic psychiatry's responsibility for increasing the recognition of mood disorders and risk of suicide in primary care. *Current Opinion in Psychiatry, 23*, 157–166. doi:10.1097/YCO.0b013e328333e195

LeVine, E. S., & Foster, E. O. (2010). Integration of psychotherapy and pharmacotherapy by prescribing medical psychologists: A psychobiosocial model of care. In R. E. McGrath & B. A. Moore (Eds.), *Pharmacotherapy for psychologists: Prescribing and collaborative roles* (pp. 105–131). Washington, DC: American Psychological Association. doi:10.1037/12167-006

LeVine, E. S., & Wiggins, J. (2010). In the private practice setting: A survey of the experiences of prescribing psychologists. In R. E. McGrath & B. A. Moore (Eds.), *Pharmacotherapy for psychologists: Prescribing and collaborative roles* (pp. 153–171). Washington, DC: American Psychological Association. doi:10.1037/12167-008

Lynch, P., & Galbraith, K. M. (2003). Panic in the emergency room. *Canadian Journal of Psychiatry/Revue canadienne de psychiatrie*, *48*, 361–366.

Maizels, M., Smitherman, T. A., & Penzien, D. B. (2006). A review of screening tools for psychiatric comorbidity in headache patients. *Headache: The Journal of Head and Face Pain, 46*(Suppl. 3), 98–109. doi:10.1111/j.1526-4610.2006.00561.x

Marceau, L. D., Link, C., Jamison, R. N., & Carolan, S. (2007). Electronic diaries as a tool to improve pain management: Is there any evidence? *Pain Medicine, 8*, S101–S109. doi:10.1111/j.1526-4637.2007.00374.x

Mayfield, D., McLeod, G., & Hall, P. (1974). The CAGE questionnaire: Validation of a new alcoholism instrument. *American Journal of Psychiatry, 131*, 1121–1123. doi:10.1176/appi.ajp.131.10.1121

McGrady, A. (2007). Psychophysiological mechanisms of stress: A foundation for the stress management therapies. In P. M. Lehrer, R. L. Woolfolk, & W. E. Sime (Eds.), *Principles and practice of stress management* (3rd ed., pp. 16–37). New York, NY: Guilford Press.

McGrath, R. E. (2010). Prescriptive authority for psychologists. *Annual Review of Clinical Psychology, 6*, 21–47. doi:10.1146/annurev-clinpsy-090209-151448

Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain, 1*, 277–299. doi:10.1016/0304-3959(75)90044-5

Melzack, R. (1987). The short-form McGill Pain Questionnaire. *Pain, 30*, 191–197. doi:10.1016/0304-3959(87)91074-8

Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people for change* (2nd ed.). New York, NY: Guilford Press.

Millon, T., Antoni, M., Millon, C., Minor, S., & Grossman, S. (2006). *Millon Behavioral Medicine Diagnostic: MBMD manual* (2nd ed.). Minneapolis, MN: Pearson.

Mitchell, A. J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research, 43*, 411–431. doi:10.1016/j.jpsychires.2008.04.014

Mori, D. L., Gallagher, P., & Milne, J. (2000). The Structured Interview for Renal Transplantation—SIRT. *Psychosomatics*, *41*, 393–406. doi:10.1176/appi.psy.41.5.393

Mpofu, E., & Oakland, T. (Eds.). (2010). *Rehabilitation and health assessment: Applying ICF guidelines*. New York, NY: Springer.

Nair, A. K., Gavett, B. E., Damman, M., Dekker, W., Green, R. C., Mandel, A., . . . Stern, R. A. (2010). Clock Drawing Test ratings by dementia specialists: Interrater reliability and diagnostic accuracy. *Journal of Neuropsychiatry and Clinical Neurosciences*, *22*, 85–92. doi:10.1176/appi.neuropsych.22.1.85

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*, 695–699. doi:10.1111/j.1532-5415.2005.53221.x

Olbrisch, M. E., Benedict, S. M., Ashe, K., & Levenson, J. L. (2002). Psychological assessment and care of organ transplant patients. *Journal of Consulting and Clinical Psychology, 70*, 771–783. doi:10.1037/0022-006X.70.3.771

Olbrisch, M. E., Levenson, J. L., & Hamer, R. (1989). The PACT: A rating scale for the study of clinical decision-making in psychosocial screening of organ transplant candidates. *Clinical Transplantation, 3*, 164–169.

Olsson, E. M. G., El Alaoui, S., Carlberg, B., Carlbring, P., & Ghaderi, A. (2010). Internet-based biofeedback-assisted relaxation training in the treatment of hypertension: A pilot study. *Applied Psychophysiology and Biofeedback, 35*, 163–170. doi:10.1007/s10484-009-9126-x

Parker, C., & Philp, I. (2004). Screening for cognitive impairment among older people in Black and minority ethnic groups. *Age and Ageing, 33*, 447–452. doi:10.1093/ageing/afh135

Pearson, S. D., Katzelnick, D. J., Simon, G. E., Manning, W. G., Helstad, C. P., & Henk, H. J. (1999). Depression among high utilizers of medical care. *Journal of General Internal Medicine, 14*, 461–468. doi:10.1046/j.1525-1497.1999.06278.x

Peek, C. J. (2003). A primer of biofeedback instrumentation. In M. S. Schwartz & F. Andrasik (Eds.), *Biofeedback: A practitioner's guide* (3rd ed., pp. 43–87). New York, NY: Guilford Press.

Piotrowski, C. (2007). Review of the psychological literature on assessment instruments used with pain patients. *North American Journal of Psychology*, *9*, 303–306.

Provenzano, D. A., Fanciullo, G. J., Jamison, R. N., McHugo, G. J., & Baird, J. C. (2007). Computer assessment and diagnostic classification of chronic pain patients. *Pain Medicine, 8*, S167–S175. doi:10.1111/j.1526-4637.2007.00379.x

Radloff, L. S. (1977). The CES–D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. doi:10.1177/014662167700100306

Roth-Roemer, S., Kurpius, S. R., & Carmin, C. (Eds.). (1998). *The emerging role of counseling psychology in health care*. New York, NY: Norton.

Roy-Byrne, P. P., Russo, J., Cowley, D. S., & Katon, W. J. (2003). Panic disorder in public sector primary care: Clinical characteristics and illness severity compared with "mainstream" primary care panic disorder. *Depression and Anxiety, 17*, 51–57. doi:10.1002/da.10082

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption—II. *Addiction, 88*, 791–804. doi:10.1111/j.1360-0443.1993.tb02093.x

Sedlacek, K., & Taub, E. (1996). Biofeedback treatment of Raynaud's disease. *Professional Psychology: Research and Practice, 27*, 548–553. doi:10.1037/0735-7028.27.6.548

Snowden, L. R., & Pingitore, D. (2002). Frequency and scope of mental health service delivery to African Americans in primary care. *Mental Health Services Research, 4*, 123–130. doi:10.1023/A:1019709728333

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder—The GAD-7. *Archives of Internal Medicine*, *166*, 1092–1097. doi:10.1001/archinte.166.10.1092

Spitzer, R. L., Kroenke, K., & Williams, J. B. W., & the Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA*, *282*, 1737–1744. doi:10.1001/jama.282.18.1737

Stolee, P., Hillier, L. M., Esbaugh, J., Bol, N., McKellar, L., & Gauthier, N. (2005). Instruments for the assessment of pain in older persons with cognitive impairment. *Journal of the American Geriatrics Society, 53*, 319–326. doi:10.1111/j.1532-5415.2005.53121.x

Tallman, B., Shaw, K., Schultz, J., & Altmaier, E. (2010). Well-being and posttraumatic growth in unrelated donor marrow transplant survivors: A nine-year longitudinal study. *Rehabilitation Psychology, 55*, 204–210. doi:10.1037/a0019541

Tariq, S. H., Tumosa, N., Chibnall, J. T., Perry, M. H., III, & Morley, J. E. (2006). Comparison of the Saint Louis University Mental Status Examination and the Mini-Mental State Examination for detecting dementia and mild neurocognitive disorder—A pilot study. *American Journal of Geriatric Psychiatry, 14*, 900–910. doi:10.1097/01.JGP.0000221510.33817.86

Tasmuth, T., von Smitten, K., Hietanen, P., Kataja, M., & Kalso, E. (1995). Pain and other symptoms after different treatment modalities of breast cancer. *Annals of Oncology, 6*, 453–459.

Tedeschi, R. G., & Calhoun, L. G. (1995). *Trauma and transformation: Growing in the aftermath of suffering*. Thousand Oaks, CA: Sage.

Tedeschi, R. G., & Calhoun, L. G. (1996). The Posttraumatic Growth Inventory: Measuring the positive legacy of trauma. *Journal of Traumatic Stress, 9*, 455–471. doi:10.1007/BF02103658; doi:10.1002/jts.2490090305

Tomich, P. L., & Helgeson, V. S. (2004). Is finding something good in the bad always good? Benefit finding among women with breast cancer. *Health Psychology, 23*, 16–23. doi:10.1037/0278-6133.23.1.16

Turk, D. C., & Melzack, R. (Eds.). (2001). *Handbook of pain assessment* (2nd ed.). New York, NY: Guilford Press.

Twillman, R. K., Manetto, C., Wellisch, D. K., & Wolcott, D. L. (1993). The Transplant Evaluation Rating Scale: A revision of the psychosocial levels system for evaluating organ transplant candidates. *Psychosomatics*, *34*, 144–153. doi:10.1016/S0033-3182(93)71905-2

Watson, L. C., & Pignone, M. P. (2003). Screening accuracy for late-life depression in primary care: A systematic review. *Journal of Family Practice*, *52*, 956–964.

Wheat, A. L., & Larkin, K. T. (2010). Biofeedback of heart rate variability and related physiology: A critical review. *Applied Psychophysiology and Biofeedback, 35*, 229–242. doi:10.1007/s10484-010-9133-y

Wong, C. L., Holroyd-Leduc, J., Simel, D. L., & Straus, S. E. (2010). Does this patient have delirium? Value of bedside instruments. *JAMA, 304*, 779–786. doi:10.1001/jama.2010.1182

Woody, R. H. (2009). Ethical considerations of multiple roles in forensic services. *Ethics and Behavior, 19*, 79–87. doi:10.1080/10508420802623690

Zaro, J. S., Batchelor, W. F., Ginsberg, M. R., & Pallak, M. S. (1982). Psychology and the JCAH: Reflections on a decade of struggle. *American Psychologist, 37*, 1342–1349. doi:10.1037/0003-066X.37.12.1342

CHAPTER 18

OUTCOMES ASSESSMENT IN HEALTH SETTINGS

Mark E. Maruish

The interest in and necessity for outcomes measurement and accountability in the era of health care reform provides a unique opportunity for psychologists to use their training and skills in assessment (Maruish, 2002, 2004). However, the extent to which psychologists and other trained professionals become key and successful contributors to any outcomes initiative, or use outcomes assessment procedures solely for their own purposes, will depend on their understanding of what "outcomes" are, important aspects of outcomes measurement, and what is involved in the application of outcome information.

The purpose of this chapter is to provide an overview of important aspects of outcomes assessment in health care settings, particularly with regard to the why, what, how, and when of outcomes assessment as well as the analysis of outcomes data. Because the primary audience for this handbook includes psychological practitioners, researchers, and students, the focus of this chapter is the assessment of behavioral health-related outcomes in either behavioral health care settings or, as is becoming more common, in general medical settings.

WHAT ARE OUTCOMES?

Before discussing outcomes assessment, it is important to have a clear understanding of what is meant by the term *outcomes*. Outcomes are probably best understood as just one component of quality of care. Donabedian (1985) identified three dimensions of quality of care. The first is structure, or the various aspects of the organization providing the care (e.g., staffing, physical facilities). The second dimension is process, which refers to the specific types of services provided to a given patient (or group of patients) during a specific episode of care. The third dimension of quality of care, outcomes, refers to the results of the specific treatment that was rendered. These results can include any number of variables that are relevant to stakeholders in the patient's care. As Sederer, Dickey, and Eisen (1997) noted,

> Outcome for patients, families, employers, and payers is not simply confined to symptomatic change. Equally important to those affected by the care rendered is the patient's capacity to function within a family, community, or work environment or to exist independently, without undue burden to the family and social welfare system. Also important is the patient's ability to show improvement in any concurrent medical and psychiatric disorder. . . . Finally, not only do patients seek symptomatic improvement, but also they want to experience a subjective sense of health and well being. (p. 2)

Portions of this chapter are reproduced or adapted with permission from *Psychological Testing in the Age of Managed Behavioral Health Care*, by M. E. Maruish, 2002, Mahwah, NJ: Erlbaum, copyright 2002 by Erlbaum; and *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment: Vol. 1. General Considerations* (3rd ed., "Introduction," pp. 1–64, and "Implementation of a Behavioral Health Outcomes Program," pp. 215–272), by M. E. Maruish (Ed.), 2004, Mahwah, NJ: Erlbaum, copyright 2004 by Erlbaum.

DOI: 10.1037/14048-018
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

It is important to recognize that the outcomes, or results, of treatment should not imply a change in only a single aspect of functioning. Thus, the term is used commonly in plural form (i.e., outcomes) to convey that interventions typically affect or produce changes in multiple aspects of the patient's life (Berman, Rosen, Hurt, & Kolarz, 1998).

STATUS OF OUTCOMES ASSESSMENT IN HEALTH CARE SETTINGS

Assessment of health care outcomes has taken place in general medical and behavioral health care settings in one form or another for decades. Repetition of tests and procedures during a course of treatment represents a means of monitoring and assessing the outcomes of that treatment. However, it is only within the past couple of decades that outcomes assessment conducted as a formal or standardized procedure by individual clinicians or health care organizations has started to become an integral part of the way health care is delivered.

Behavioral Health Care Settings

Since the 1990s, the behavioral health care field has witnessed accelerating growth in the level of interest and development of outcomes assessment programs. A 1998 study conducted by the Center for Mental Health Services and reported by Manderscheid, Henderson, and Brown (2001) found that 85% of 676 outpatient mental health facilities had systems in place to measure adult outcomes. In 1999, Levy Merrick, Garnick, Horgan, and Hodgkin (2002) conducted a national survey of 434 managed care organizations (MCOs). They found that almost half of the MCO products (48.9%) conducted behavioral health outcomes assessments.

A few surveys have focused specifically on outcomes assessment conducted by psychologists. In a survey conducted in 1995 by the American Psychological Association's (APA's) Committee for the Advancement of Professional Practice, Phelps, Eisman, and Kohut (1998) found that assessment was the second most prevalent activity of their sample, occupying an average of 16% of the professional time of the nearly 16,000 respondents. They also found that 29% of the respondents were involved in outcomes assessment, with the highest rate of use of outcomes measures (40%) reported by psychologists in medical settings. To investigate the use of outcomes measures in clinical practice, Hatfield and Ogles (2004, 2007) conducted a survey of 2,000 licensed psychologists who were randomly drawn from APA members who had paid APA's special practice assessment fee. Among the 874 respondents, 37.1% reported that they used some outcomes assessment in their practice. Of these, 74.4% indicated that they used patient self-report measures, and 61.2% reported using clinician-completed measures.

In a survey of reported usage of 45 child and adolescent assessment instruments, Cashel (2002) found that 20% of the respondents indicated that they formally assessed treatment outcomes either "frequently" or "routinely," whereas another 48% indicated they did so "sometimes." This figure compares with an earlier survey by Bickman et al. (2000), in which 54% of clinicians treating child and adolescent patients reported using outcomes measures.

General Medical Settings

The enactment of the Mental Health Parity Act of 1996, the industry's realization of the benefits of one-stop health care, accreditation standards, the growing belief that potential long-term health care cost savings can result from the appropriate treatment of behavioral disorders, and other circumstances served as the impetus for a more pervasive integration of primary and behavioral health care services throughout the United States. Goldstein et al. (2000) summarized the state of these affairs by noting that a significant percentage of primary care patients experience significant psychological symptomatology or distress. The value the behavioral health care professional brings to the primary care setting is attested to daily in primary care practices throughout the country.

One of the most significant contributions that psychologists can make to the integration of medical and behavioral health care is through the establishment and use of psychological assessment services. Information obtained from psychometrically sound self-report measures and other instruments (e.g.,

clinician rating scales, parent-completed instruments) can assist the primary care provider in several types of clinical decision-making activities, including screening for the presence of mental health or substance abuse problems, planning a course of treatment, and monitoring patient progress. In addition, testing can be used to measure the outcomes of treatment that has been provided to patients with mental health or substance abuse problems, thus assisting in determining what works for whom. Moreover, a psychologist's expertise can extend to the use of other, medically oriented tests and surveys that provide information (e.g., pain level, role functioning) useful with patients suffering solely from physical problems (e.g., diabetes, arthritis).

Beyond the primary care setting, medical populations for which psychological assessment is useful is quite varied. As Todd (1999) observed,

> Today, it is difficult to find any organization in the healthcare industry that isn't in some way involved in disease management. . . . This concept has quickly evolved from a marketing strategy of the pharmaceutical industry to an entrenched discipline among many managed care organizations. (p. xi)

Where can outcomes assessment fit into these programs? Health plans recognize the value that psychological assessment can bring to their programs, including the ability to help identify and track medical patients with comorbid behavioral health problems. Diabetes, asthma, and other medical disorders are accompanied frequently by depression and anxiety that can significantly affect the patient's quality of life, morbidity, and, in some cases, mortality. Early identification and treatment of comorbid behavioral health problems in patients with chronic medical diseases, along with an assessment of outcomes of that treatment can dramatically affect the course of the disease and the toll it takes on the patient.

Additional information about psychological assessment in medical settings is provided in Chapter 17 of this volume. Also, Chapter 13 in this volume provides a more detailed discussion of the role that psychological assessment can play in treatment.

SOURCES OF OUTCOMES DATA

One of the most important considerations related to how outcomes data are obtained is from where or whom these data should come. Certain types of outcomes data will necessitate the use of specific sources of information, whereas other types can be obtained legitimately from more than one source. The type of setting and population also may have a bearing on the selection of the best source of data (Berman, Hurt, & Heiss, 1996).

Patient Self-Report

In many cases, the most important data will be those obtained directly from the patient using self-report instruments. Indeed, self-report measures appear to be the most commonly used sources of outcomes information in behavioral health care (Farnsworth, Hess, & Lambert, 2001). It is important to note, however, that using patient self-report data may be viewed with suspicion by some (Strupp, 1996). This author has personally witnessed the rejection of outcomes information that contradicted staff impressions, purely because it was based on patient self-report data. The implication was that such data are not valid. Generally, such concerns are not justified.

Indeed, only relatively recently has the physical (medical) health-related field of research become more accepting of patient self-report outcomes data. Whereas the objective data obtained from sources such as medical claims, lab results, and electronic imaging (e.g., magnetic resonance images; X-rays; positron emission tomography [PET] scans) once served as the primary source of outcomes information for health services research, clinical trials and the like, one now finds increasing acceptance and use of patient-reported outcomes, or self-report measures, in these areas of research.

Collateral Sources

Other types of data gathering tools may be substituted for self-report measures. Rating scales completed by the clinician or other members of the treatment staff may provide information that is as useful as that elicited directly from the patient. In those cases in which the patient is severely disturbed, unable to give valid and reliable answers

(e.g., younger children), unable to read, or is an otherwise inappropriate candidate for a self-report measure, clinical rating scales can serve as a valuable substitute for gathering information about the patient. Related to these instruments are parent-completed inventories for child and adolescent patients. These are particularly useful in obtaining information about the behavior of children or adolescents that otherwise might not be known. Information also might be obtained from other patient collaterals (e.g., spouses, teachers, employers), all of whom may offer valuable information by themselves or in combination with other information.

Administrative Data

Another potential source of outcomes information is administrative data. In many of the larger organizations, this information can be retrieved easily through an organization's claims and authorization databases, data repositories and warehouses, and other databases that make up the organization's management information system. Data related to the patient's diagnosis, dose and regimen of medication, physical findings, and other types of data typically stored in these systems can be useful in evaluating the outcomes of therapeutic intervention. Medical records also may provide this and other important diagnostic and treatment-related information (e.g., physical findings, diagnosis).

Multiple Sources

Many would agree that the ideal approach for gathering outcomes data would be to use multiple sources (Berman et al., 1998; Bieber, Wroblewski, & Barber, 1999; Strupp, 1996). Inherent in this approach, however, are increased burden and costs and the potential for contradictory information and concomitant questions about how to proceed when contradictions occur. Thus, one must be prepared with approaches for resolving contradictory information that make sense from the perspective of all parties concerned with the treatment of the patient.

PURPOSE OF OUTCOMES ASSESSMENT

Just as it is important to be clear about what is meant by "outcomes," it is equally important to clarify the three general purposes for which outcomes assessment may be used. The first purpose of assessing outcomes is outcomes *measurement*. This process involves nothing more than pre- and posttreatment assessment of one or more variables to determine the amount of change that has occurred in the specified variable(s) during the episode of care. This approach is limited in its utility compared with other approaches to outcomes assessment.

A more useful approach is that of outcomes *monitoring*. Outcomes monitoring involves tracking changes in the status of one or more outcomes variables at multiple points in time. Assuming a baseline assessment at the beginning of treatment, reassessment may occur one or more times during the course of treatment (e.g., weekly, every two sessions), at the time of termination, during one or more periods of posttermination follow-up, or at two or more of these points in time. Whereas treatment progress monitoring is used to determine deviation from the expected course of improvement, outcomes monitoring focuses on revealing aspects of the therapeutic process that seem to affect change.

The third, and most useful, purpose of outcomes assessment is that of outcomes *management*. Dorwart (1996) has defined outcomes management as "the use of monitoring information in the management of patients to improve both the clinical and administrative processes for delivering care" (pp. 46–47). Whereas Dorwart appears to view outcomes management as relevant to the individual patient during an episode of care, this author views it as a means to improve the quality of services offered to the patient populations served by the provider, not to any one patient. Information gained through the assessment of patients can provide the organization with indications of what works best for whom and under what set of circumstances, thus helping to improve the quality of services for all patients.

In the outcomes assessment process, a baseline measurement is taken. In some cases, this step may be followed by the treatment monitoring process discussed earlier. Frequently, the patient is assessed at the termination of treatment, although assessment at the end of treatment may not always be the case. Posttreatment follow-up measurement may occur, with or without measurement at the time of

termination. Follow-up may involve more than one remeasurement at various points in time. Commonly used intervals include 3 months, 6 months, and 12 months posttermination. Again, the results from those participating in follow-up remeasurement are usually combined with those of other individuals for group data analysis. The information gleaned from this analysis gives the psychologist or organization a sense of what has worked with whom and can be used to guide treatment practices with other patients in the future.

WHAT TO MEASURE

As discussed earlier, treatment may affect multiple facets of a patient's life. Thus, it is not a simple matter to determine exactly which outcomes should be measured. The specific aspects or dimensions of patient functioning that are selected for measurement will depend on the purpose for which the assessment is conducted. Sometimes the needs or interests of the various stakeholders in the treatment of the patient drive the types of outcomes that are measured. These interests may vary greatly, as was suggested by Sederer et al. (1996) at the beginning of this chapter.

In behavioral health settings, probably the most frequently measured variable is symptomatology or psychological or mental health status because disruption in this dimension is often the most common reason why people seek behavioral healthcare services in the first place. However, there are other reasons for seeking help. Common examples include difficulties in coping with various types of life transitions (e.g., a new job, a recent marriage or divorce), or an inability to deal with the behavior of others (e.g., spouse, coworkers), or general dissatisfaction with life. Thus, one may find that for some patients, improved functioning on the job, at school or with family or friends is much more relevant and important than symptom reduction. For other patients, improved quality of life or sense of well-being may be more meaningful.

In medical settings, amelioration of symptomatology, cure or control of disease (e.g., diabetes, asthma), and patient engagement in preventative health behaviors (e.g., inoculation against the flu, maintenance of a healthy weight) are common goals or outcomes. However, with the ongoing integration of behavioral health services in primary care and other health care settings, some of the same outcomes variables that are important in behavioral health settings—especially those pertaining to quality of life—also are relevant here.

Symptomatology

In evaluating psychiatric symptomatology, one may consider the patient's general level of distress or disturbance or the presence or level of one or more specific types of symptomatology. General level of psychological or emotional distress is often assessed using multiscale instruments (e.g., Minnesota Multiphasic Personality Inventory—2 [MMPI–2]; Butcher et al., 2001), some of which enable the user to combine results from several or all scales into single measures of psychological distress (e.g., the Global Severity Index of the Symptom Checklist-90-R; Derogatis, 1994).

There are numerous narrowly focused symptom-specific scales appropriate for use in general medical or behavioral health care settings, such as the Zung and Beck Depression and Anxiety scales. Measures of depression and anxiety are specifically mentioned here because, in the vast majority of instances, depression and anxiety are psychological symptoms or problems that both behavioral health and medical patients are most likely to present. If one must measure something in a medical or behavioral health setting, he or she is on pretty safe ground measuring either or both of these symptoms. However, other outcomes variables may be just as or even more important.

General Health Status

During the past 2 decades, there has been an increasing interest in the measurement of health status in both behavioral and general medical health care delivery systems. Initially, this interest was shown primarily within those organizations and settings focused on the treatment of physical diseases and disorders. In recent years, behavioral health care organizations along with psychologists and other behavioral health care providers have recognized the value of evaluating the patient's general level of health. It is important to recognize that the term *health* means more than just the absence of disease or debility; it also implies a state of well-being

throughout the individual's physical, psychological, and social spheres of existence (World Health Organization, as cited in Stewart & Ware, 1992).

Measures of general health status and functioning are appropriate for use across both patient and "healthy" populations. Probably the most widely used and respected generic health status measure is the SF-36v2 Health Survey (SF-36v2; Maruish, 2011). The SF-36v2 measures eight domains of health—four addressing mental health-related constructs and four addressing physical health-related constructs—that reflect the World Health Organization's concept of health.

Subjective Well-Being

Concomitant with society's relatively recent focus on maintaining good health and preventing disease and illness, in both the physical and psychological realms, has been the identification of subjective well-being as an important aspect of psychological health, one that should be attended to and promoted, and, consequently, one that should be measured. Well-being is one of those nebulous constructs that are difficult to define or describe, although most people have a good sense of what it is. Frankish, Herbert, Milsum, and Peters (1999) defined the term in a manner that is consistent with or important in a psychotherapeutic context:

> While well-being is associated with health, a consensus is emerging that the term well-being implies a wider emphasis than does health on the individual's sense of wholeness, in all its physical, mental, emotional, social and spiritual aspects. It expresses the individual's capacity to cope with stress without losing effective functioning. (pp. 41–42)

It is difficult to identify examples of well-known or widely used instruments that were developed to measure psychological well-being. Typically, well-being is equated with quality of life (QOL), and QOL measures are ones that are commonly cited (discussed in the immediately following section). Indeed, the relationship between the two constructs can be a bit unclear. As with the case of measures of role functioning (discussed later), this author is most familiar with sets of a few Likert-type rating items that tap into one's feeling of contentment with life and their ability to cope.

Quality of Life

Andrews, Peters, and Teesson (1994) indicate that most definitions of QOL describe a multidimensional construct encompassing physical, affective, cognitive, social, and economic domains. Seid, Varni, and Jacobs (2000) indicated other distinctions in the QOL arena. One has to do with the differences between QOL and health-related quality of life (HRQOL). As Seid et al. noted,

> Quality of life encompasses all aspects of an individual's life, including housing, environment, work, school, a safe neighborhood, and the like, which are traditionally beyond the scope of the healthcare system. HRQOL refers specifically to those domains of an individual's health that are potentially within the influence of the healthcare system. (p. 18)

Why measure QOL? Walters (2009) identified several reasons. In addition to providing insight into issues other than symptoms, which traditionally have been the focus of treatment and outcomes measurement, QOL assessment provides a means of facilitating communication with patients and finding out more of the extent of the problems that patients experience. QOL information also can help establish a patient's views and preferences. Moreover, Walters indicated that it may have prognostic value and can be used to help make population-level treatment decisions. Measurement of HRQOL is important in both behavioral and general health care settings; however, it is particularly relevant in nonbehavioral health care settings.

Similar to the case with health status measures, the other distinction that Andrews et al. (1994) made is between generic and condition-specific measures of QOL. Generic measures are designed to assess aspects of life that are generally relevant to most people; condition-specific measures are focused on aspects of the lives of particular disease or disorder populations. A more extensive discussion of

generic and condition-specific outcomes measures in general is presented in a later section of this chapter.

There are several instruments available that specifically measure for QOL. One example is the Quality of Life Inventory (Frisch, 1994). However, measures of other constructs are frequently used or referred to as sources of information about QOL. For example, one will frequently see references to the SF-36v2 as being an indicator of HRQOL. Thus, one should be discerning when investigating potential measures of QOL or HRQOL.

Role Functioning

Role functioning is viewed as an important variable to address in the course of assessing the effect of a physical or mental disorder on an individual's life functioning. How the person's ability to work, perform daily tasks, or interact with others is affected by a disease or disorder is important to consider in developing a treatment plan and monitoring progress over time. One such aspect relates to satisfactory performance on the job or, in the case of children or college students, at school. Improvement in one's marital relationship in general as well as sexual performance, parenting skills, and other identified marital or family problem areas can all be important indicators of positive treatment-related changes. Improvement in relationships outside of the family (e.g., with friends, coworkers) can also be a good indicator of psychological improvement.

There are several measures of role functioning that are currently available, including the Katz Adjustment Scales (KAS; Katz & Warren, 1997). However, the KAS and similar instruments generally tend to be too lengthy for use by individual providers or health care systems. Often, one will find the use of a few very general items, either as part of an intake assessment or as a component of an outcomes measurement system, to be sufficient. In other instances, general health status measures (e.g., SF-36v2) that incorporate scales tapping into broad aspects of role functioning will be used to help assess this domain.

Pain

Pain may be one of the most important variables to assess in almost any general or specialty medical care setting, as pain is often the reason people seek medical care in the first place. A patient's perception of pain is also a key outcomes variable in behavioral health settings that offer specialized units or programs for coping with pain, regardless of its origin (i.e., psychological vs. organic). One should not be surprised to see improvement of perceived pain to covary with the improvement of other outcomes variables. For example, the amelioration of pain may lead to improvement in role functioning and in both general and health-related QOL.

Substance Use

Aside from individuals being treated for addiction and other substance abuse problems, the assessment of alcohol and other substance use can serve as one indicator of the degree to which an individual relies on inappropriate or otherwise maladaptive mechanisms to cope with the challenges of daily living. Like pain, changes in the report of the amount or frequency of use of alcohol or other substances will often covary with changes reported on other outcomes variables, such as pain. Here, however, one may be just as likely to see deterioration on other outcomes variables with improvement in substance abuse behavior, as in a case when reduction or cessation of alcohol or other drugs leads to a report of increased levels of pain or generalized anxiety, which the patient had previously been using drugs to cope with or control. This type of information would indicate that treatment has focused more on a symptom than the underlying problem.

Condition-Specific Variables

In most general medical and behavioral health settings, the assessment of every patient treated using a common set of outcomes variables is desirable. There are times, however, that some of the most important and useful outcomes that should be measured are those specific to the disease, disorder, or condition with which the patient presents. Assessment of these condition-specific outcomes in addition to or instead of the set of generic or common outcomes variables can, therefore, be very important. For example, in a large hospital offering several inpatient medical specialty services, level of cognitive functioning may be an important outcomes variable

for the hospital's stroke unit, whereas on the psychiatry unit, level of depression is more salient.

HOW TO MEASURE

Once the decision of *what* to measure has been made, one must then decide *how* it should be measured. In some cases, the "what" will dictate the "how." In others, there will be multiple options for the how of the measurement.

Types of Instruments

Standardized versus not standardized. Standardized instruments are always preferred over nonstandardized instruments. Standardization refers to "*uniformity of procedure* in administering and scoring [a] test [or survey]" (Anastasi & Urbina, 1997 p. 6). Urbina (2004) added that "the purpose of standardizing test procedures is to make all of the variables that are under the control of the examiner as uniform as possible, so that everyone who takes the test will be taking it the same way" (p. 2). In addition, both Anastasi and Urbina have pointed to the development and use of norms as a major component of standardization.

Self-report versus other report. One of the most important considerations related to how outcomes data are obtained is from where or whom these data should come. Certain types of outcomes data will necessitate the use of specific sources of information, whereas other types can be obtained from more than one source. In addition, the type of setting and population will have a bearing on the selection of the best source of data (Berman et al., 1996). The issue of self-report versus other report was addressed earlier in this chapter. Also, the reader is referred to Chapter 11 of this volume for a more complete discussion of the assessment personality and psychopathology with self-report inventories. Moreover, Volume 1, Chapter 19, this handbook, provides a general discussion of objective personality testing.

Generic versus condition-specific. Broad measures of physical and mental health have the advantage of not being limited to the examination of one particular type of disease, condition, or psychopathology. In behavioral health settings, using these types of measures is a significant advantage when one wants to measure the outcomes of treatment on a population that presents with a wide range of psychopathology and problems spanning the entire range of severity. Similarly, in general health care settings where QOL is a common patient-reported outcome variable, generic measures enable the assessment and comparison of individuals presenting with a broad continuum of health states—including healthy people—on a common metric, with results reported as a profile of scores or a single index of health (Fayers & Machin, 2007; Hays, 2005; Walters, 2009).

A major drawback of generic instruments is that they are likely to be longer than symptom-specific instruments in that they attempt to measure the patient's level of distress on multiple symptom domains. Another drawback may occur in certain settings (e.g.,specialty clinics or even other general treatment settings) where only one or two types of patients are usually seen. Conversely, brevity is likely to be the biggest advantage of symptom-, disease-, or condition-specific measures, as is their limited yet relevant symptom focus in settings that usually treat only one type of presenting problem. At the same time, brevity may limit a measure's reliability, and the narrow focus may result in a failure to detect and measure other significant problems or their improvement resulting from intervention.

Unidimensional versus multidimensional. By definition, unidimensional instruments assess only one type of disorder or symptom domain and thus are of limited utility. They are most useful in situations in which only a single symptom domain or disorder is of interest, such as in clinical drug trials or treatment programs focused on the alleviation of only one type of symptomatology (e.g., depression, migraine headaches). Most are brief and are generally used when one wants to screen for or monitor a particular type of symptomatology, pathology, or functional impairment. Good examples are the Beck Depression Inventory (2nd ed.; BDI-II; Beck, Steer, & Brown, 1996) and the Headache Impact Test (HIT-6; Bayliss & Batenhorst, 2002).

As alluded to earlier, multidimensional instruments can serve a variety of purposes that facilitate

therapeutic interventions. They may be used on initial contact with the patient to screen for the need for service and, at the same time, yield information that is useful for diagnosis and treatment planning. Indeed, some such instruments (e.g., MMPI–2, SF-36v2) may make available supplementary, content-related, or other special scales or indices that can assist in addressing specific treatment considerations (e.g., motivation to engage in treatment, predicted medical expenditures). Other multiscale instruments might be useful in identifying specific problems that may be unrelated to the patient's chief complaints (e.g., low self-esteem). Use of such instruments in a pre- and posttreatment fashion can provide information related to the outcomes of a patient's treatment on multiple dimensions or domains.

Criteria for Selection of Outcomes Measures

Availability of instrumentation for outcomes assessment purposes is not an issue. However, selection of the appropriate instrument(s) for outcomes assessment is a matter requiring careful consideration. Inattention to an instrument's intended use, the demonstrated psychometric characteristics associated with its intended use, its limitations, and other aspects related to its practical application can result in invalid or otherwise useless outcomes data, not to mention misguided treatment and potentially harmful consequences for a patient.

Regardless of the type of measure one might consider to use for outcomes assessment, psychologists frequently must choose between many product offerings. Table 18.1 presents a summary of criteria and associated considerations that are recommended for the selection of outcomes assessment instruments. Some of these may seem obvious, but one would be surprised how easily some of these considerations can be overlooked. Reliability and test validity are among the most important considerations in the choice of psychological measurement for any purpose and are discussed in Volume 1, Chapters 2 and 4, this handbook, respectively.

It is also important to recognize that one will probably not find a single outcomes measure that would meet the needs of all stakeholders in the care

TABLE 18.1

Criteria for Selecting Outcomes Assessment Instruments

Criteria	Important considerations
Brevity	Is considered short from the patient's perspective
Reading level	Requires no higher than an eighth-grade reading level, with sixth grade or lower preferable, or can be administered through another mode that does not require reading (e.g., live interview, IVR) and yields comparable results
Psychometric integrity	Meets generally accepted standards for validity and reliability Has demonstrated responsiveness (for individual data) or sensitivity (for group data) to changes in patient status
Relevancy to the intended purpose of the assessment	Is appropriate for measuring the targeted outcomes domain(s) in the targeted population
Availability of relevant normative data	Has norms that are appropriate for the targeted population
Cost	Inexpensive to use for multiple administrations to a single patient
Acceptability to patients	Does not include questions that patients are likely to find to be unnecessary, embarrassing, or not face valid
Ease of use	Is easy to administer, score, and interpret
Comprehensibility of results to all parties involved in treatment	Results can be easily understood by the provider, patient, family members, and other relevant stakeholders
Actionable information	Provides direction about how to improve the quality of services and what, if any, changes need to be made in treatment
Overall practicality or feasibility	Given all considerations, is practical for use in the intended setting, with the intended population, for the intended purpose(s)

of a single patient or a patient population (Eisen & Dickey, 1996; Norquist, 2002). Given this possibility, Ogles, Lambert, and Fields (2002) have offered a few suggestions that should be heeded in selecting outcomes instrumentation. First, one should know the trade-offs. Identifying the pros and cons of the instrumentation in terms of meeting one's personal or organizational needs will allow one to make an informed decision. Second, one should know the audience. Instrument selection should always take into consideration who will be the end users of the obtained information. When there are multiple audiences for this information (as is often the case), the needs of each must be balanced. Third, one also needs to recognize resource limitations. It is important to assess the burden that a given instrument will pose on the organization's or practitioner's financial and human resources. For an instrument under consideration, one might ask: Is the information the instrument yields worth the cost of obtaining it?

Modes and Technologies for Outcomes Assessment

Perhaps the most common means of gathering outcomes information for individual patients is through the administration of a paper-and-pencil version of the instrument or through a carefully scripted, standardized face-to-face interview while the patient is in the clinician's office. However, as has always been the case, someone has had the foresight to develop applications of current technological advances that are used every day to the practice of psychological assessment. Just as at one time the personal computer held the power of facilitating the in-office assessment process, the Internet, fax, and interactive voice response (IVR) technologies have been developed to make the assessment process easier, quicker, and more cost-effective.

Internet. An Internet-based outcomes assessment process is straightforward. The clinician accesses the website on which the desired instrumentation resides. The desired measure is selected for administration, and then the patient completes the test online. The data are scored and entered into the website's database, and a report is generated and transmitted to the clinician or patient or both through the Internet. Turnaround time for receiving the report is usually only a matter of minutes. In addition to outcomes assessment purposes, the archived data can be used later for any of a number of purposes, including treatment monitoring, regularly scheduled reporting of aggregated data, psychometric test development, and other statistical purposes.

Faxback. The process for implementing faxback technology also is fairly simple. A specially developed, test-specific paper-and-pencil answer sheet is completed by the patient and then faxed in to a central facility where the data are both entered into a database and then scored. In those systems in which several tests are available, the answer sheet for a given test contains numbers or other types of code that tell the scoring and reporting software which test is being submitted. A report is generated and faxed or e-mailed to the clinician or made available to the clinician or patient or both on a secure website within a few minutes. Later, the stored data can be used in the same ways as those gathered by an Internet-based system.

IVR. One of the more recent applications of new technology to the administration, scoring, and reporting of results of psychological tests can be found in the use of IVR systems. In general, IVR technology allows for the gathering of information using a telephone. Its applicability to test administration, data processing, and data storage is simple. Survey administration through IVR typically involves the presentation of prerecorded instructions and survey questions to which patients respond orally (on systems utilizing voice recognition software) or, more commonly, by using the telephone keypad to select a numbered, multiple-choice response option or to give a numeric response, such as that pertaining to age or the frequency of a behavior or event.

A summary of the advantages and disadvantages of these and other common modes of outcomes assessment is presented in Table 18.2.

WHEN TO CONDUCT OUTCOMES ASSESSMENTS

An important issue for individual clinicians and health care organizations wanting to integrate

TABLE 18.2

Advantages and Disadvantages of Outcomes Assessment Modalities

Assessment modalities	Advantages	Disadvantages
Mail-out/mail-back	Does not require any special equipment or software Good for research involving large groups, over a large geographic area, or repeated administration over time Enables assessment of enduring effects of treatment long after treatment termination	Cannot be used with patients with limited or no reading ability Lack of control of testing environment Costs for postage and follow-up Determining the most effective survey method May require costly follow-up to obtain data
Interview	Does not require any special equipment or software Provides a test administration solution for patients with limited or no reading ability May be the only way some patients will agree to provide the outcomes information being sought	May require interviewer training, including associated time and cost Requires clinician or staff time, including associated cost
Internet	Immediate access to updated or enhanced versions of software Results immediately available for clinical decision-making Enables computer-adaptive test (CAT) administration of measures based on item–response theory (IRT)	Cannot be used with patients with limited or no reading ability Possible security issues Requires access to the Internet
Faxback	Assessment is completed in paper-and-pencil format Facilitates data entry for scoring and reporting Facilitates database entry for aggregation and analysis of sample or population data	Cannot be used with patients with limited or no reading ability Possible security issues May require patient access to fax machine
IVR technology	No additional equipment required for patient administration Available for patient use 24 hours/day, 7 days/week Provides a test administration solution for patients with limited or no reading ability	Possible security issues Administration must be initiated by the patient May require costly follow-up to obtain data

outcomes assessment into their standard way of delivering services is deciding *when* the two or more assessments of each patient receiving treatment should take place. Although a seemingly simple matter to address, it requires careful consideration before arriving at a decision that may have significant implications later on.

General Considerations

There are no hard and fast rules or widely accepted conventions related to when outcomes should be measured. The common practice is to assess the patient on the selected outcomes variables at least at treatment initiation, and then again at termination or discharge. The problem with relying solely on the "pre-post" assessment approach is that sometimes the time at which treatment ends is unpredictable (Lyons, Howard, O'Mahoney, & Lish, 1997), making self-report data difficult to obtain. However, there are solutions to this problem.

Measurement can take place at other points in time; that is, during treatment and on postdischarge follow-up. However, this approach still raises the issue of when to assess outcomes. Should it be done on the basis of the number of sessions that have been completed (e.g., every third session), or at specific time points from the date of treatment initiation (e.g., every 4th week)? Also, how many times should a patient be asked to complete an outcomes protocol? With regard to the first issue, Berman et al. (1996) argued for the "time-from-initial-contact model" over the "session model," as this model

allows for meaningful data gathering in settings offering multiple levels of care and at posttreatment when there is no session that can be used to gauge the next time of measurement once treatment has been terminated.

As for the number of times one imposes on the patient to complete the outcomes measures, one solution Lyons et al. (1997) offered is to incorporate the outcomes protocol into the routine assessment activities that normally take place during the course of treatment. This approach would have the effect of repeated assessment being perceived as standard practice for the clinician or organization, not as something extra the patient is asked to do.

Instrument-Related Factors Determining Frequency of Assessment

In addition to the considerations just discussed, decisions about when to readminister an outcomes assessment instrument must take into account aspects of the instrumentation being used. Many tests and surveys used for outcomes assessment purposes include items that ask the respondent to consider a specific time interval when responding to the question. For example: "During the past 4 weeks, how often have you . . ." or "During the past 24 hours, how many times have you . . . ?" In cases such as these, readministration of the instrument should not take place any sooner than the amount of time the respondent is asked to consider has elapsed since the last administration. Readministration of the instrument any sooner results in overlapping reporting periods for the measured outcomes variables and the meaning of the results of both periods may become muddied to the point of uselessness.

Situational Factors Determining Frequency of Assessment

Sederer et al. (1997) suggested that clinicians or organizations should take into account some very important considerations about when to conduct outcomes measurement. One particularly important consideration touched on by these authors is the minimum amount of time that one would expect for an intervention to begin to have an effect on the variable(s) of interest. For example, reporting outcomes data for patients receiving outpatient substance abuse treatment on a weekly basis may not allow enough time to plan and implement an intervention and allow that intervention to have an effect and show results during the next reporting period. Also, one might not expect to see improvements in functioning until later in treatment. According to the phase model of psychotherapy (Howard, Lueger, Maling, & Martinovich, 1993), mental health improvement is evidenced first by improvement subjective well-being, then symptom relief, which is then followed by functional improvement; however, this sequence of improvement is not always the case (e.g., see Bryan, Morrow, & Appolonio, 2009).

ANALYSIS OF OUTCOMES DATA

Decisions about how one plans to analyze outcomes data can have a significant impact on many of the considerations discussed earlier. Not having a decision about one's analytic strategy before implementing either an individual or organizationwide outcomes initiative can have disastrous consequences later on (see Dawson, Doll, Fitzpatrick, Jenkinson, & Carr, 2010). The questions that outcomes data are intended to answer should drive the types of analyses to be performed. In turn, knowing what types of analyses need to be conducted may have a significant bearing on what data are collected, how they are collected, and when they are collected.

Analysis of Individual Patient Data

There are two general approaches to the analysis of outcomes data for determining if a patient has changed on one or more outcomes variables from one point in time to another. The first is by determining whether changes in patient scores on outcomes measures are statistically significant. The other is by establishing whether these changes are clinically significant.

The issue of clinically significant change has received a great deal of attention in psychotherapy research over the past few decades. This focus is due, at least in part, to the work of Jacobson and his colleagues (e.g., Jacobson, Follette, & Revenstorf,

1984; Jacobson & Truax, 1991) and others (e.g., Christensen & Mendoza, 1986; Speer, 1992), which came at a time when researchers began to recognize that traditional statistical comparisons do not reveal a great deal about the efficacy of therapy.

Jacobson and Truax (1991) broadly defined the clinical significance of treatment as "its ability to meet standards of efficacy set by consumers, clinicians, and researchers" (p. 12). Furthermore, they noted a lack of consensus about what these standards should be. From their perspective, clinically significant change could be conceptualized in one of several ways. However, for them, for clinically significant change to have occurred, the measured level of functioning following the therapeutic episode would have to be closer to the mean of the functional population than to that of the dysfunctional population. Jacobson and Truax considered this approach to be the least arbitrary of the approaches, and they provided different recommendations for determining cutoffs for clinically significant change, depending on the availability of normative data.

At the same time, these same investigators noted the importance of considering the change in the measured variables of interest from pre- to posttreatment in addition to the patient's functional status at the end of therapy. Accordingly, Jacobson et al. (1984) proposed the concomitant use of a reliable change index (RCI) to determine whether change is clinically significant. This index, later modified on the recommendation of Christensen and Mendoza (1986), is the pretest score minus the posttest score divided by the standard error of the difference of the two scores. Following the research of Lambert, Hansen, and Finch (2001), those who begin treatment in a functional or "normal" range but make reliable change in the direction of improvement should be considered improved but not clinically significantly improved. Both those who begin treatment in the functional range and deteriorate into the dysfunctional range as well as those who begin treatment in the dysfunctional range and deteriorate further would be considered deteriorators. Additional discussion concerning the evaluation of pre- to posttreatment changes can be found in Chapter 13 in this volume.

Related to the RCI is the *responder criterion*, which may be more commonly used in general medical health care outcomes assessment. According to Maruish (2011), the RCI approach of Jacobson and his colleagues appears to be overly conservative because it assumes that the baseline and follow-up scores are uncorrelated. Furthermore, whereas a 95% confidence interval (equivalent to a 5% significance level) is used as a standard in group-level analyses, this criterion may be overly conservative for analyses of individual patients, where the risk of falsely identifying change must be balanced against the risk of overlooking true change. Thus, Maruish proposed that it is more reasonable to assume a baseline-to-follow-up correlation (e.g., .10) while using a less conservative confidence interval (e.g., 80%).

There are other approaches to analyzing individual patient data for clinically significant change. Excellent discussions of the RCI and some of these other methods can be found in Hsu (1999), Kazdin (1999), and Maruish (2011). Interested readers are encouraged to review these and other publications on the topic before deciding which approach is best for them.

Analysis of Group Aggregated Data

Changes in groups of patients from one point in time to another typically have been examined through the use of any of a number of tests of statistical differences in mean scores. Generally, this method is quite appropriate and not likely to draw much criticism (assuming that the most appropriate statistical test has been used). Although it may be important to know that a real change in a sample or population has taken place, these types of analyses do not provide any indication of the magnitude of that change.

One means of determining whether change in group-level scores from one time point to another is meaningful or important is through the comparison of observed score differences to a minimally important difference (MID) value established for a given scale or index. This approach is frequently used in HRQOL research (e.g., see Norman, Sloan, & Wyrwich, 2003) and in the analysis of clinical trial data

to support patient-reported outcomes claims to regulatory agencies such as the U.S. Food and Drug Administration (e.g., see Revicki, Hays, Cella, & Sloan, 2008). Essentially, change equal to or surpassing MID value represents a change that patients themselves view as meaningful rather than a change that a clinician or researcher identifies as meaningful from a clinical perspective (Maruish, 2011). MIDs can be established using either of two approaches. In the *anchor-based approach*, the anchor or criterion represents a clinical marker or health-related event and the MID score value for the scale or measure in question represents an important (i.e., nontrivial) change in that criterion. The *distribution-based approach* is based on the distribution of scores for the scale or measure, with interpretation of the findings with respect to the relationship between the size of the difference (i.e., within- or between-group differences) and some measure of variability (e.g., standard deviation). Interested readers are referred to Fayers and Machin (2007), Revicki et al. (2008), Walters (2009), and Maruish (2011) for further information and discussion.

To answer questions related to the magnitude or importance of change, many health care researchers use statistics to measure effect size (ES). ES can be defined as "an interpretation of the size of the observed effect . . . in terms of the variability among individuals" (Osoba & King, 2005, p. 249). As Fayers and Machin (2007) have pointed out, the term is applied to several standardized measures of change; however, the ES statistic that is frequently utilized is computed by dividing the difference between the pre- and posttreatment means by the pretreatment standard deviation ($ES = [m_1 - m_2] / s_1$). Cohen (1988) interpreted ES values of less than 0.2 as indicating a trivial or no effect; values between 0.2 and 0.5 indicate a small effect; values between 0.5 and 0.8 suggest a moderate effect; and values greater than 0.8 indicate a large effect. Note that others advocate for different cutoffs for determining the magnitude of effects (e.g., see Hopkins, 1997). Regardless, as Kazis, Anderson, and Meenan (1989) pointed out, ESs provide for a more interpretable measure of change and allow for comparison of differences on different measures within or between outcomes systems.

Another approach would be to analyze the data using both ES and methods of significance testing. This approach is being seen more frequently in the published literature. Doing so would not require significantly more effort beyond that for one or the other method, but it would better satisfy the needs of all stakeholders and other interested parties.

A "better-same-worse" analysis represents a simple yet informative means to analyze group outcomes data. This approach involves the comparison of the percentages of those determined to have gotten better, remained the same, and gotten worse on a selected outcomes variable, from one assessment point to another. Categorizing a patient as being better, the same, or worse on the variable of interest at one point in time relative to another is based on whether the difference in the scores for the variable meets a predetermined criterion. This change criterion can be based on any of several statistics (e.g., standard error of measurement, standard deviation, MID) for the outcomes measure (see Maruish & Kosinski, 2009). For example, Martin et al. (2007) investigated changes from baseline in Physical Component Summary (PCS) and Mental Component Summary (MCS) scores from the SF-8 Health Survey (Ware, Kosinski, Dewey, & Gandek, 2001) separately, for groups of coronary artery disease and heart failure patients, separately, participating in disease management programs, on a quarterly basis for a period of 1 year. With a 95% confidence interval as the criterion, patients whose PCS or MCS scores were equal to or greater than the baseline score by 1.96 *SEM* were considered to have improved or to be "better"; patients whose PCS or MCS scores were equal to or less than the baseline score by −1.96 *SEM* were considered to have deteriorated or to be "worse"; and the remainder of the patients were considered to be the "same." Chi-square tests for significant differences in the category membership percentages were also performed each quarter.

Another common approach to analyzing group outcomes data is to compare the results with some standard. Taking the route of comparing outcomes assessment results against some standard begs the question of which standard to use. Even before that, however, one must decide which type of standard

best meets the needs of his or her outcomes assessment efforts. There are a few options here, each with its own set of advantages and drawbacks.

First, population-specific data can serve as a standard against which to compare performance. Unlike benchmark or industry standards, this approach relies on data that are more specific to and representative of different types of populations and the characteristics that distinguish them from other populations. Standardized normative data that typically accompany published psychological tests is a good example. These data permit a fair comparison of groups of patients with like groups of patients or nonpatients, thus eliminating some of the potential effects of confounding variables. When population-specific comparison data are not available for the outcomes variables that are important to the stakeholders in the patient's treatment, risk adjustment procedures are frequently used to allow fair comparisons among different groups of patients.

Second, published data sets such as HEDIS (National Committee for Quality Assurance, 2004) can provide valuable information about the success other organizations have achieved on standard performance measures. Use of this information in this way is referred to as *benchmarking,* "an ongoing process of comparing [an] organization's performance on services, practices, and outcomes with some identified standard, such as . . . competitors' performance" (Christner, 1997, p. 2). Benchmarking allows the clinician, health care organizations, and other stakeholders to see how the clinician or organization fares in comparison with similar clinicians or organizations. The downside is that performance measures on which industrywide data are available may not always be what the organization or its stakeholders feel are most important or relevant to the care of their patients.

The third standard is that set by the clinician or organization itself. To some degree, it probably will be based on a combination of what the industry standard is and what the organization sees as being realistic given the people it serves, the resources available, expectations from stakeholders, accreditation and regulatory requirements, and whatever other demands it must meet to remain successful and solvent.

CONSIDERATIONS FOR REPORTING OUTCOMES ASSESSMENT DATA

An important but often neglected aspect of outcomes assessment is how the findings and related data will be reported. Reporting is generally addressed in Chapter 3 of this volume, but there are there are a few considerations that require particular attention when developing reports of outcomes assessment findings. The first is what the intent of the report is. This intention should be a relatively easy decision if one has taken the time to define the purpose of the outcomes assessment endeavor and what questions it is supposed to answer. Of course, there may be multiple reasons for assessing outcomes and multiple questions that need to be answered, and trying to address all questions and matters of interest may be problematic from a reporting standpoint. The amount of available information also may be problematic. The issue then becomes one of determining (a) what information is considered primary, secondary, and so on, and (b) how much of that information can be presented and remain meaningful.

Just as important as the intent of the report is to whom the report will be directed. Often, these two factors go hand in hand. Many stakeholders in the patient's care may want to receive a report of patient progress. Problems may arise when the needs of more than one party must be met. One solution to this problem is to develop different reports for the different stakeholders, each of which includes only the information that each party needs or wants.

Finally, some of the same situational and test-specific considerations discussed earlier in the section titled When to Measure will affect the reporting of outcomes findings. Also, one or more third parties (e.g., payers, accreditation or licensing bodies) may dictate the reporting cycle, as may costs associated with reporting outcomes (e.g., materials, equipment, manpower) and, thus, may play a part in how frequently individuals are assessed and reports are generated.

CONCLUSIONS

The movement toward measuring treatment outcomes in medical and behavioral healthcare settings

has been gradually gathering steam over the past 2 decades. Spurred by a need to assess the effect of treatment, to identify what works for whom, to justify payment for services and, overall, to control spiraling health care costs, outcomes assessment is more and more establishing itself as a routine part of health care. It is an area of health services in which psychologists and others with psychometric expertise can make important contributions as the U.S. health care system begins to undergo significant changes.

Many factors must be considered before an individual health care provider or a health care system begins to assess outcomes in any formal or routine way. These factors have to do with the *why, what, how,* and *when* of assessment. The *why* has to do with the purpose the outcomes data will serve or the questions the data will answer. It goes beyond the general purposes of measurement, monitoring, or management of outcomes to more specific reasons. These reasons could range from determining what type of treatment is best for certain types of patients or patients with specific problems; to providing evidence for demonstrating to a potential patient or health care plan that the services offered by the individual provider, health care practice, or organization are effective; to meeting a requirement for accreditation or licensing from an external body.

The *why* of outcomes assessment must be clearly established before deciding on *what* to assess. Sometimes, the choice of *what* is simple. For instance, in behavior health care settings, the primary focus is likely to be on symptom resolution, so assessment of symptom intensity and frequency may be at the forefront. Elsewhere, the options for what to assess can be numerous, depending on the setting. In general medical settings providing a wide range of medical services, one may have to opt for assessing a common outcome that allows comparisons of treatment effectiveness across departments while at the same time meeting the needs of several medical disciplines (e.g., HRQOL). Individual medical or behavioral disciplines can, of course, supplement this type of measure with additional measures that are more relevant or specific to the types of patients that they treat.

After deciding *what* to assess, one must then determine *how* to obtain the desired information. This determination not only involves the type of instrument to be used (e.g., generic vs. condition-specific, unidimensional vs. multidimensional) but also the manner in which it is to be administered to the patient or a collateral (e.g., parent, spouse). Aside from the traditional paper-and-pencil mode of administration, several technologies that can facilitate the efficient collection of outcomes data (e.g., fax, IVR, Internet) are currently available.

In addition to the selection of the type of assessment instrument and technology to employ for collecting and processing outcomes data, the *how* also involves the manner in which outcomes data will be analyzed. This matter is an often overlooked yet very important aspect in determining *how* the desired outcomes data will be collected. It is particularly important in outcomes programs where the focus is on analysis and reporting of group-level data. For example, knowing ahead of time what type of statistical analyses need to be performed to yield meaningful and useful information from the outcomes data can have a direct bearing on (a) the type of instrumentation that should be used (e.g., use of an instrument that yields continuous vs. categorical data), (b) the amount of patient and staff burden that will be required to obtain the size of the sample that is needed to adequately power the analyses, and (c) the type of assessment technology that will be needed to facilitate the data collection and analysis efforts.

The final question is one of *when* to assess. There are no hard and fast rules guiding *when* outcomes data should be collected from patients. This decision will depend on several factors, such as the aspects of the outcomes instrumentation itself, aspects of the particular patient population being served (including typical disease course and expected length of treatment, particularly in multispecialty medical practices), the burden placed on patients and staff in collecting the data, and internal and external demands for outcomes information.

In closing, it is important to recognize that all decisions about outcomes assessment should be guided by *practicality*. First, one must always consider the availability and commitment of resources—both financial and manpower—that would be required to implement a system of

outcomes assessment. Second, like the burden that outcomes data collection places on treatment and support staff, the burden imposed on patients must also be weighed. Asking patients to complete outcomes instruments that are lengthy or asking them to complete any instrument too many times will increase patient burden and resistance and, consequently, decrease the likelihood of obtaining much useable data. Third, the degree to which any system of outcomes assessment is successful will depend on the buy-in of the setting's personnel. This includes not only the in-the-trenches staff but also the middle and upper levels of management, especially in larger health care settings and systems. If upper management does not actively support the outcomes initiative, it is unlikely that others will be committed to the endeavor. Finally, outcomes data should yield actionable information; that is, information that can help guide decision making or service delivery in general, regardless of whether the beneficiary of those services are the patients who provide the data or future patients presenting with similar problems.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Andrews, G., Peters, L., & Teesson, M. (1994). *The measurement of consumer outcomes in mental health.* Canberra, Australia: Australian Government Publishing Service.

Bayliss, M. S., & Batenhorst, A. S. (2002). *The HIT-6: A user's guide.* Lincoln, RI: QualityMetric.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Pearson.

Berman, W. H., Hurt, S. W., & Heiss, G. E. (1996). Outcomes assessment in behavioral healthcare. In C. E. Stout, J. Oher, & G. A. Theis (Eds.), *The complete guide to managed behavioral care* (pp. II-D.1–II-D.10). New York, NY: Wiley.

Berman, W. H., Rosen, C. S., Hurt, S. W., & Kolarz, C. M. (1998). Toto, we're not in Kansas anymore: Measuring and using outcomes in behavioral health care. *Clinical Psychology: Science and Practice, 5*, 115–133. doi:10.1111/j.1468-2850.1998.tb00139.x

Bickman, L., Rosof-Williams, J., Salzer, M. S., Summerfelt, W. T., Noser, K., Wilson, S. J., & Karver, M. S. (2000). What information do clinicians value for monitoring adolescent client progress and outcomes? *Professional Psychology: Research and Practice, 31*, 70–74. doi:10.1037/0735-7028.31.1.70

Bieber, J., Wroblewski, J. M., & Barber, C. A. (1999). Design and implementation of an outcomes management system within inpatient and outpatient behavioral health settings. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 171–210). Mahwah, NJ: Lawrence Erlbaum Associates.

Bryan, C. J., Morrow, C., & Appolonio, K. K. (2009). Impact of behavioral health consultant interventions on patient symptoms and functioning in an integrated family medicine clinic. *Journal of Clinical Psychology, 65*, 281–293. doi:10.1002/jclp.20539

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory—2: Manual for administration, scoring, and interpretation* (rev. ed.). Minneapolis: University of Minnesota Press.

Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice, 33*, 446–453. doi:10.1037/0735-7028.33.5.446

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index [Letter to the editor]. *Behavior Therapy*, *17*, 305–308. doi:10.1016/S0005-7894(86)80060-0

Christner, A. M. (1997, January). Using baselines and benchmarks can sharpen your outcomes evaluation. *Behavioral Health Outcomes*, *2*, 1–3.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C., & Carr, A. J. (2010). Routine use of patient reported outcome measures in healthcare settings. *British Medical Journal, 340*, 464–467. doi:10.1136/bmj.c186

Derogatis, L. R. (1994). *SCL-90-R Symptom Checklist-90-R: Administration, scoring, and procedures manual* (3rd ed.). Minneapolis, MN: NCS Pearson.

Donabedian, A. (1985). *Explorations in quality assessment and monitoring: The methods and findings in quality assessment: An illustrated analysis* (Vol. 3). Ann Arbor, MI: Health Administration Press.

Dorwart, R. A. (1996). Outcomes management strategies in mental health: Applications and implications for clinical practice. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 45–54). Baltimore, MD: Williams & Wilkins.

Eisen, S. V., & Dickey, B. (1996). Mental health outcome assessment: The new agenda. *Psychotherapy:*

Theory, Research, Practice, Training, 33, 181–189. doi:10.1037/0033-3204.33.2.181

Farnsworth, J. R., Hess, J. Z., & Lambert, M. J. (2001, August). *Frequency of outcomes measures used in psychotherapy*. Poster presented at the 109th Annual Convention of the American Psychological Association, San Francisco, CA.

Fayers, P. M., & Machin, D. (2007). *Quality of life: The assessment, analysis, and interpretation of patient-reported outcomes* (2nd ed.). Chichester, England: Wiley.

Frankish, C. J., Herbert, C., Milsum, J. H., & Peters, H. F. (1999). Measurements of positive health and well-being. In G. C. Hyner, K. W. Peterson, J. W. Travis, J. E. Dewey, J. J. Foerster, & E. M. Framer (Eds.), *SPM handbook of health assessment tools* (pp. 41–48). Pittsburgh, PA: The Society of Prospective Medicine & The Institute for Health and Productivity Management.

Frisch, M. B. (1994). *Manual and treatment guide for the Quality of Life Inventory*. Minneapolis, MN: National Computer Systems.

Goldstein, L., Bershadsky, B., & Maruish, M. E. (2000). The INOVA primary behavioral health care pilot project. In M. E. Maruish (Ed.), *Handbook of psychological testing in primary care settings* (pp. 735–760). Mahwah, NJ: Erlbaum.

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice, 35*, 485–491. doi:10.1037/0735-7028.35.5.485

Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcomes measures and others do not. *Administration Policy in Mental Health and Mental Health Services Research, 34*, 283–291. doi:10.1007/s10488-006-0110-y

Hays, R. D. (2005). Generic versus disease-targeted instruments. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials* (2nd ed., pp. 3–8). New York, NY: Oxford University Press.

Hopkins, W. G. (1997). *A new view of statistics: A scale of magnitude for effect sizes*. Retrieved from http://www.sportsci.org/resource/stats/efffectmag.html

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting and Clinical Psychology, 61*, 678–685. doi:10.1037/0022-006X.61.4.678

Hsu, L. M. (1999). Caveats concerning comparisons of change rates obtained with five models of identifying significant client changes: Comment on Speer and Greenbaum (1995). *Journal of Consulting and Clinical Psychology, 67*, 594–598. doi:10.1037/0022-006X.67.4.594

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. doi:10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19. doi:10.1037/0022-006X.59.1.12

Katz, M. M., & Warren, W. L. (1997). *Katz Adjustment Scales Relative Report Form (KAS-R) manual*. Los Angeles, CA: Western Psychological Services.

Kazdin, A. E. (1999). The meaning and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332–339. doi:10.1037/0022-006X.67.3.332

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. *Medical Care, 27*, S178–S189. doi:10.1097/00005650-198903001-00015

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172. doi:10.1037/0022-006X.69.2.159

Levy Merrick, E., Garnick, D. W., Horgan, C. M., & Hodgkin, D. (2002). Quality measurement and accountability for substance abuse and mental health services in managed care organizations. *Medical Care, 40*, 1238–1248. doi:10.1097/00005650-200212000-00010

Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. D. (1997). *The measurement and management of clinical outcomes in mental health*. New York, NY: Wiley.

Manderscheid, R. W., Henderson, M. J., & Brown, D. Y. (2001). Status of national accountability efforts at the millennium (Publication No. SMA 01-3537). In R. W. Manderscheid & M. J. Henderson (Eds.), *Mental health, United States, 2000. Section 2: Decision support 2000+* (pp. 1–10). Washington, DC: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration.

Martin, M., Blaisdell-Gross, B., Fortin, E. W., Maruish, M. E., Manocchia, M., Sun, X., . . . Ware, J. E., Jr. (2007). Health-related quality of life of heart failure and coronary artery disease patients improved during participation in disease management programs: A longitudinal observational study. *Disease Management*, *10*, 164–178. doi:10.1089/dis.2007.103612

Maruish, M. E. (2002). *Psychological testing in the age of managed behavioral health care*. Mahwah, NJ: Erlbaum.

Maruish, M. E. (2004). Introduction. In M. E. Maruish (Ed.), *The use of psychological testing for treatment*

planning and outcomes assessment: Vol. 1. General considerations (3rd ed., pp. 1–64). Mahwah, NJ: Erlbaum.

Maruish, M. E. (Ed.). (2011). *User's manual for the SF-36v2 Health Survey* (3rd ed.). Lincoln, RI: QualityMetric.

Maruish, M. E., & Kosinski, M. (2009). *A guide to the development of certified Short Form survey interpretation and reporting capabilities*. Lincoln, RI: QualityMetric.

National Committee for Quality Assurance. (2004). *HEDIS 2004: Vol. 6. Specifications for the Medicare Health Outcomes Survey*. Washington, DC: Author.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life. The remarkable universality of half a standard deviation. *Medical Care, 41*, 582–592. doi:10.1097/01.MLR.0000062554.74615.4C

Norquist, G. S. (2002). Role of outcome measurement in psychiatry. In W. W. Ishak, T. Burt, & L. I. Sederer (Eds.), *Outcome measurement in psychiatry: A critical review* (pp. 3–13). Washington, DC: American Psychiatric Publishing.

Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcomes assessment*. New York, NY: Wiley.

Osoba, D., & King, M. (2005). Meaningful differences. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 243–257). Oxford, England: Oxford University Press.

Phelps, R., Eisman, E. J., & Kohut, J. (1998). Psychological practice and managed care: Results of the CAPP practitioner survey. *Professional Psychology: Research and Practice, 29*, 31–36. doi:10.1037/0735-7028.29.1.31

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*, 102–109. doi:10.1016/j.jclinepi.2007.03.012

Sederer, L. I., Dickey, B., & Eisen, S. V. (1997). Assessing outcomes in clinical practice. *Psychiatric Quarterly, 68*, 311–325. doi:10.1023/A:1025490930088

Sederer, L. I., Dickey, B., & Hermann, R. C. (1996). The imperative of outcomes assessment in psychiatry. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 1–7). Baltimore, MD: Williams & Wilkins.

Seid, M., Varni, J. W., & Jacobs, J. R. (2000). Pediatric health-related quality-of-life measurement technology: Intersections between science, managed care, and clinical care. *Journal of Clinical Psychology in Medical Settings, 7*, 17–27. doi:10.1023/A:1009541218764

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991). revisited. *Journal of Consulting and Clinical Psychology, 60*, 402–408. doi:10.1037/0022-006X.60.3.402

Stewart, A. L., & Ware, J. E., Jr. (1992). *Measuring functioning and well-being*. Durham, NC: Duke University Press.

Strupp, H. H. (1996). The tripartite model and the Consumer Reports study. *American Psychologist, 51*, 1017–1024. doi:10.1037/0003-066X.51.10.1017

Todd, W. E. (1999). Introduction: Fulfilling the promise of disease management: Where are we today? Where are we headed? In S. Heffner (Ed.), *Disease management sourcebook 2000: Resources and strategies for program design and implementation* (pp. xi–xxiii). New York, NY: Faulkner & Gray.

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: Wiley.

Walters, S. J. (2009). *Quality of life outcomes in clinical trials and health-care evaluation: A practical guide to analysis and interpretation*. Chichester, United Kingdom: Wiley.

Ware, J. E., Jr., Kosinski, M., Dewey, J. E., & Gandek, B. (2001). *How to score and interpret single-item health status measures: A manual for users of the SF-8 Health Survey*. Lincoln, RI: QualityMetric.

PART III

COUNSELING PSYCHOLOGY

CHAPTER 19

ASSESSMENTS OF INTERESTS

Bryan J. Dik and Patrick J. Rottinghaus

The construct of interests is one of the most commonly assessed in psychology and has continuously ranked among the most vigorously studied and applied individual differences constructs for the past century (Savickas & Spokane, 1999). Interests have been studied by cognitive psychologists approaching the construct as an emotion (Silvia, 2006) and sporadically studied by industrial–organizational psychologists (Dawis, 1992), but the construct has been so heavily and rigorously investigated within counseling psychology that it is recognized as a cornerstone of the field (Betsworth & Fouad, 1997). Interests also are the most frequently assessed construct in career counseling practice (Hansen, 2005). Along with abilities and values, interests are one of the "Big Three" constructs considered central to career choice and development applications (Swanson & D'Achiardi, 2005).

This chapter provides an overview of the assessment of interests in psychology. We contextualize interest assessment by summarizing its history and the most salient theoretical approaches to understanding interests currently. A brief overview of the major domains of research on the construct follows, including the structure of interests, the relation of interests to other individual differences dimensions, interest stability, and differences across sex and cultural groups. We describe the uses and methods of interest assessment including techniques used to construct interest inventories, then provide an overview of four popular interest inventories. Finally, we provide recommendations for using interest inventories in counseling practice.

HISTORY OF INTEREST ASSESSMENT

Before the 1920s, vocational guidance professionals did not have useful assessment tools and had to rely on the use of interviews and client self-study methods to help increase their clients' self-knowledge of interests and other relevant characteristics such as abilities, needs, values, and personality. Although Alfred Binet's intelligence test first demonstrated (in 1908) that individual differences could be reliably measured, Frank Parsons's (1909/2005) model of person–environment fit (P-E fit) most often is cited as the impetus for investigating the role of interests in career choice (Donnay, 1997). Parsons's deceptively simple model held that wise career choices involved understanding (a) work-related aspects of the self ("person") and (b) different occupations in the world of work ("environment"), and then (c) using "true reasoning" to find an optimal match between the person and the available environments ("fit"). Interests were viewed as a key component of a person's work personality. Attempts to measure interests date to E. L. Thorndike's 1912 study of rank-ordered interests among college students and a questionnaire designed by T. L. Kelley in 1914 (Campbell, 1971). However, the most direct antecedent of the modern interest inventory was a 1919 seminar conducted by C. S. Yoakum at the Carnegie Institute of Technology (now Carnegie-Mellon University) in Pittsburgh (Campbell, 1971). In this seminar, Yoakum and students developed a pool of approximately 1,000 items that formed the basis of the first generation of interest inventories. One of the inventories that evolved from this item pool was

DOI: 10.1037/14048-019
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

the Strong Vocational Interest Blank (SVIB) developed by E. K. Strong Jr.

Strong, who had earlier studied under Thorndike and James McKeen Cattell at Columbia University, was head of the Bureau of Educational Research at Carnegie during the time of Yoakum's seminar. After his move in 1923 to Stanford University, Strong guided Karl Cowdery, a graduate student, to use Freyd's (1923) Occupational Interest Inventory (another offspring of Yoakum's item pool) to differentiate members of particular occupations—engineers, lawyers, and physicians—on the basis of their item responses. This project inspired the development of the more refined SVIB (Strong, 1927). Strong constructed the SVIB using the empirical method of contrast groups, in which items retained for the each of the 10 occupational scales were those that differentiated the likes and dislikes of members of a particular occupation from those of a group of men in general. (The first SVIB for women was published in 1933.) The SVIB has since been revised and expanded a half-dozen times and, now called the Strong Interest Inventory (SII; Donnay, Morris, Schaubhut, & Thompson, 2005), is one of the most widely used of all psychological inventories in research and practice (Walsh & Betz, 1995). As further evidence of the effect of the Strong, several other popular interest inventories have been developed by individuals who previously were involved in scale construction research for the Strong, such as David Campbell's Campbell Interest and Skill Survey (CISS; Campbell, Hyne & Nilsen, 1992) and Charles Johansson's Career Assessment Inventory (Johansson, 2003).

Not long after the SVIB was published, G. F. Kuder, a psychometrician known for his contributions to reliability theory, introduced an inventory that measured interests using content-related homogenous dimensions. The Kuder Preference Record (Kuder, 1939), predecessor to what became the Kuder Occupational Interest Survey (KOIS; Kuder & Zytowski, 1991) and is now the Kuder Career Search with Person Match (KCS; Zytowski, 2009), was scored on a rational basis and provided ipsative (i.e., intraindividual) scores. Kuder also used Cleman's lambda to provide an index of the similarity of an individual respondent to an occupational reference group. The contrasting psychometric approaches of Strong and Kuder have been described as the foundation of contemporary vocational interest measurement (Donnay, 1997). Indeed, the interest inventories available today owe much to these pioneers.

VOCATIONAL INTERESTS IN THEORETICAL CONTEXT

In psychology, interests have been conceptualized in two ways: (a) "interest" as a transient emotional state and (b) "interests" as stable, enduring dispositions (Dik & Hansen, 2008). The study of interest as an emotion has largely been the domain of cognitive psychologists. They note that "interest," synonymous with curiosity or inquisitiveness, meets all the major criteria typically used to classify a construct as an emotion. First, interest is marked by distinct expressive signals, such as stillness of the head, parted lips, widened eyelids, and increased eye contact (e.g., Reeve, 1993; Reeve & Nix, 1997; Wallerstein, 1954), even among infants (e.g., Langsdorf, Izard, Rayias, & Hembree, 1983). Second, interest is accompanied by unique, salient, and coherent subjective feelings, both in terms of self-report and behavioral criteria (e.g., comprehensiveness, depth, and length of time engaging a stimulus; Langsdorf et al., 1983; Reeve, 1993; Reeve & Nix, 1997). Third, interest plays an adaptive developmental role for people in that it broadens and builds life experiences that can be helpful when facing future events (e.g., Fredrickson, 1998). Finally, interest plays a key role in personality processes by catalyzing the development of enduring, traitlike "interests" (Silvia, 2006).

This last point is salient because it bridges "interest" and "interests." That is, theoretically, affective interest leads to repeated encounters with an activity or idea through a variety of pathways, resulting in the encoding of those encounters into cognitive scripts, which form the basis of dispositional interests. In addition to this "interest-and-interests" model (Silvia, 2001, 2006), other theories of how stable interests develop propose a central role for need fulfillment (e.g., Deci, 1992), self-justification (e.g., Weick, 1964), and self-observation generalizations and task approach skills (e.g., Mitchell &

Krumboltz, 1996). Among the major theories of career choice and development, Social Cognitive Career Theory (SCCT; Lent, Brown & Hackett, 1994) provides a framework for how interests materialize. Drawing from Bandura's (1986) social–cognitive theory, SCCT postulates that people develop self-efficacy beliefs and outcome expectations for activities in which they engage; these beliefs and expectations are based on past performance, relevant feedback, vicarious learning, social encouragement, and current affective states. The theory suggests that individuals develop stable interests in activities for which they gain a sense of personal competence and which they expect will lead consistently to valued outcomes. These theoretical approaches emphasize nurture mechanisms; behavioral genetics evidence also suggests that 36% of the differences between people in interests can be accounted for by differences in their genes (Betsworth et al., 1994).

Stable, dispositional interests primarily have been examined by researchers in vocational psychology and others who investigate individual differences. Conceptualized as traitlike preferences, interests have been defined as "motivations that determine life decisions" (Walsh, 1999, p. 373) or, more simply, what interest inventories measure. The most influential theory in history applied to interests is John L. Holland's (1959, 1997b) theory of vocational types. Holland proposed that both people and work environments can be characterized according to six broad vocational types, summarized by the acronym RIASEC: Realistic (mechanical, outdoor, athletic activities), Investigative (intellectual, scientific, research activities), Artistic (fine arts, drama, writing, music, and culinary activities), Social (teaching, counseling, and social service), Enterprising (sales, managing, law, and politics), and Conventional (detail-oriented and data management). Holland proposed that the relationships among the six types can be depicted graphically by ordering them around a hexagon, with their relative proximity representing their relative similarity. He proposed that most people gravitate toward occupations that are congruent with their primary types and that the degree of fit between the types of the person and of the occupation predicts outcomes such as stability, achievement, and satisfaction. It is hard to overstate the effect of Holland's theory on research and practice related to interests; his hexagon and RIASEC typology are ubiquitous and provide the organizational structure for most commercially available interest inventories as well as several major databases of occupational information such as the Occupational Information Network (O*NET; Rounds, Smith, Hubert, Lewis, & Rivkin, 1998). Other theories of career choice and development incorporate interests indirectly or secondarily, as features of the occupational self-concept (Gottfredson, 2002; Savickas, 2002; Super, 1963) or as derivatives of skills and values (Dawis & Lofquist, 1984).

RESEARCH ON VOCATIONAL INTERESTS

Research examining vocational interests has been steady and influential to theory and practice (Savickas & Spokane, 1999). The following sections highlight areas of scholarship that are especially critical to advancing the meaning of scores on interest inventories, including structure, linkages with related constructs, stability, and differences across groups.

Structure of Interests

Research examining structural relationships between various interest domains is important to theory and practice since levels and patterns of relations between constructs affect the meaning of measured interests. Research in this area typically examines patterns of interest scores organized in a circular order or more restrictive circumplex models involving equal intervals between Holland's six types (Day & Rounds, 1998; Rounds & Tracey, 1993). Higher order levels such as sociability/conformity (Hogan, 1983) and data–ideas/people–things (Prediger, 1982), more detailed facets (e.g., math and science interests; Day & Rounds, 1997), and added dimensions (e.g., prestige; Rounds & Tracey, 1996) are addressed as well. Addressing concerns about sampling and measures used in earlier studies, recent studies have examined the cross-cultural equivalence of various theoretical structures (discussed later), largely concluding that individuals from diverse backgrounds have similar mental representations of interests (Day & Rounds, 1998;

Fouad, Harmon, & Borgen, 1997; but see Armstrong, Hubert, & Rounds, 2003).

While addressing crucial issues related to interest theory and measurement, Rounds and Tracey (1996) identified higher order levels and additional dimensions not reflected in Holland's hexagon. Their spherical model highlights orthogonal dimensions of data–ideas, people–things, and a possible third dimension—prestige. The degree of differentiation between interest dimensions is conceptualized differently depending on the respondent's degree of interest on the prestige dimension. Taken together, these dimensions can be conceptualized as a globe of occupational interests, with prestige representing the north–south axis and the RIASEC circumplex representing the equator, with interest types becoming increasingly less differentiated as one moves away from the equator. The *Personal Globe Inventory* (Tracey, 2002) is a relatively new inventory of interests and abilities that has several sets of scales reflecting the findings from research on the structure of interests. In addition to the prestige scale, results are reported at three levels to represent different levels of complexity, including (a) data, ideas, people, and things; (b) RIASEC; and (c) eight more specific areas (e.g., managing, mechanical).

Intersection Between Interests and Related Constructs

Although interests are clearly separable from abilities, values, personality, and self-efficacy, statistically and clinically significant relations between these qualities are routinely noted (Ackerman, 1996; Ackerman & Heggestad, 1997; Borgen & Lindley, 2003; Dawis, 2001; Rottinghaus & Zytowski, 2006; Spokane & Decker, 1999). Ackerman and Heggestad (1997) summarized possible patterns by noting that "abilities, interests, and personality develop in tandem, such that ability level and personality dispositions determine the probability of success in a particular task domain, and interests determine the motivation to attempt the task" (p. 239). This statement applies to Strong's (1943) often-quoted analogy of these individual differences representing a motor boat—abilities serve as a motor propelling individuals forward and interests as the rudder guiding the direction of one's attention.

Subsequent theoretical statements have mirrored research highlighting nodes of convergence between various individual differences, notably within Holland's (1959, 1997b) theory and SCCT (Lent et al., 1994). Holland conceptualized interests as the expression of one's personality. The RIASEC types are construed as amalgams of various interests, abilities, and values, operationally defined by summing interest and ability self-estimate measures in the Self-Directed Search (SDS). Decades of empirical research connects vocational interests moderately with abilities (Lubinski, 2000), ability self-estimates (Tracey, 2002), and personality (Larson, Rottinghaus, & Borgen, 2002; Sullivan & Hansen, 2004). Moreover, vocational psychologists have demonstrated direct (Rottinghaus, Betz, & Borgen, 2003; Sheu et al., 2010) and reciprocal (Nauta, Kahn, Angell, & Cantarelli, 2002; Tracey, 2002) relations between parallel measures of self-efficacy and interests.

Given the centrality of interests to vocational behavior, it is no surprise that an intricate network of relations exists. Spokane and Decker (1999) emphasized that various individual differences (e.g., interests, personality) and other aspects of the self may reflect a core latent structure. Some commonality among various groups of traits is evident, such as extraversion personality with enterprising and social interests (Ackerman, 1996; Larson et al., 2002), and between parallel measures of interest and self-efficacy (Betz & Rottinghaus, 2006). Studies examining several sets of constructs together enable researchers to identify unique contributions to various aspects of vocational behavior. For example, Rottinghaus, Lindley, Green, and Borgen (2002) demonstrated that Big Five personality traits, and parallel self-efficacy and interests for Holland RIASEC domains accounted for incremental variance in educational aspirations, increasing from 10%, 26%, and 29% with the inclusion of each set of variables, respectively.

Armstrong, Day, McVay, and Rounds (2008) found support for interests as a means of integrating abilities and personality traits into a multifaceted taxonomy of individual differences. With Ackerman's (1996; Ackerman & Heggestad, 1997) integrative model, this research supports Holland's

hexagon as an anchor for various individual differences, and offers important theoretical and practical perspectives for applied psychologists addressing work-related issues.

Stability of Interests

To the extent that interests are used as a source of information to help people make decisions about their future career, the stability of interests is a critical question. Why devote years to earning a degree in a particular academic discipline, for example, if one runs the risk of losing interest in it by the time the training period ends? A basic analysis of stability involves comparison of changes in the absolute level of measured interests in a particular domain over time; for example, a person may become significantly more interested in finance over the course of 10 years. However, the stability of interests typically is investigated by assessing a group of people at multiple points in time and calculating a correlation coefficient between time periods. This group-level approach examines the rank-order stability, or differential continuity, for those participants for a particular interest domain. Another approach to investigating stability is the individual-level "profile" approach, in which ipsative configurations (i.e., profiles) of individual participants' scores on salient interest domains are examined at two or more points in time using correlation coefficients. Using the profile of 10 activity preferences of the KOIS among 107 former high school students, Rottinghaus, Coon, Gaffey, and Zytowski (2007) reported a moderate degree of intraindividual stability (Spearman rho = .54) over the course of 30 years. Low, Yoon, Roberts, and Rounds (2005) conducted a meta-analysis of 66 longitudinal studies and reported population estimates for rank-order correlations of .60 and for profile correlations of .70. They found that population estimates for aggregated interest stability coefficients ranged from .55 to .58 from middle school through high school and then increased considerably during young adulthood (age 18+) into the .70s (even to .83 for those 25–29.9, although this estimate was based on just two studies). It remained in the .70s until dropping to .64 for participants aged 35 to 39. Longer time intervals corresponded to lower stability estimates, but no gender differences were found. Low et al. (2005) compared interest stability with that of personality traits as reported in a meta-analysis by Roberts and DelVecchio (2000). Interests were consistently more stable than personality from early adolescence through age 29; stabilities for the two constructs then converged for participants in their 30s (estimated population values for both were .62).

A third approach to examining the stability question is to consider the extent to which profiles for occupational groups may differ across the decades. Despite often seismic societal change, evidence suggests very few differences in the interest profiles of people representing particular occupations (e.g., bankers, psychologists, engineers) over periods of 40 or 50 years (Campbell, 1966; Hansen, 1988). In summary, particularly once people reach early adulthood, interests are among the most stable individual differences constructs in psychology.

Sex and Cultural Differences in Interests

The magnitude and direction of differences on interests between women and men and across cultural groups has long been a focal point of research on interests, given the relevance of such questions for the responsible use of interest inventories. Research has repeatedly found that women and men, on average, express different levels of some types of interests. Su, Rounds, and Armstrong (2009) meta-analyzed data from 47 interest inventory technical manuals, representing more than 503,000 participants, and reported effect size estimates suggesting that women report stronger artistic ($d = -.35$), social ($d = -.68$), and conventional ($d = -.33$) interests then men, who tended to report stronger realistic ($d = .84$) and investigative ($d = .26$) interests. One might expect that such differences would evaporate over time as entry into nontraditional occupations becomes more socially accepted, yet such differences persisted at both the item and scale score level for cohorts over the 50-year period between the 1930s and 1980s (Hansen, 1988). Nevertheless, Hansen (1988) found sex differences to decrease for more recent cohorts. More recently, Bubany and Hansen (2011) found in another meta-analytic study that for college student cohorts ranging from the early 1970s to mid-2000s,

there were substantial decreases in sex differences from earlier to later generations on realistic, enterprising, and conventional interests. Similarly, increases were found in enterprising interests for women, and decreases were found in realistic and investigative interests for men. Bubany and Hansen suggested these changes may echo the movement of American culture toward more egalitarian views of gender.

Research on differences across cultural groups frequently has examined the structure of interests using Holland's model (e.g., by examining the extent to which the six types conform to a circular order). Across samples of ethnically diverse groups in the United States, the similarities have been more striking than the differences (Day & Rounds, 1998; Fouad, Harmon & Borgen, 1997) in that the circular order (if not the literal equilateral hexagon) of RIASEC types is consistently supported. However, when Armstrong et al. (2003) reanalyzed data from the Day and Rounds (1998) and Fouad et al. (1997) studies using a more rigorous statistical approach (i.e., circular unidimensional scaling), they found that Holland's model may fit better for White and Asian American participants than for other groups. Fouad and Walker (2005) provided evidence that differences across racial and ethnic minority groups on interest inventory items may be due to the confounding role of a secondary trait related to culture. Several studies have examined the structure of interests in non-U.S. samples (e.g., Tracey, Watanabe, & Schneider, 1997), yielding inconclusive patterns of results. This type of research is limited by numerous factors including nonlinguistic equivalence of measures, differences in occupational opportunities, and obtaining comparable samples sufficiently large to determine multivariate structures.

Another way to examine differences across racial and ethnic groups, as Hansen (2005) reviewed, is by examining differences in the criterion-related validity of interest inventories. Evidence suggests variability across groups, from a hit rate (i.e., percentage of participants who chose a college major that corresponds to a high score on their interest profile) of 75% for Latina/Latino students to a 56% hit rate for American Indians. (As a point of comparison, the hit rate for White students is about 70%.) Of note, all of these percentages exceed the hit rate of 28% expected due to chance.

USES OF INTEREST ASSESSMENT

Most frequently, interests are assessed in an educational, counseling, or rehabilitation context to support the career exploration of individuals who are undecided about their career paths. Interest assessment can provide users with a useful heuristic (such as Holland's model) for organizing information about themselves and occupations. Interest information also can illuminate the dynamics in particular occupations (i.e., which jobs satisfy which patterns of interests), identify occupations that optimally fit an individual's interest profile, and help investigate why a current job may be dissatisfying. Interest data sometimes are combined with other sources of information in organizational settings to assist in selection, classification and placement decisions, or to facilitate team-building activities (Hansen & Dik, 2004). Interests also can assist in planning satisfying leisure activities, and in helping late-career adults plan their transitions to retirement. Finally, interest assessments can help improve relationships by objectively identifying differences between people (Campbell, Hyne & Nilsen, 1992).

METHODS OF MEASURING INTERESTS

Although quantitative measures, which provide scores representing *measured interests*, are the most commonly used approach to assessing interests, several other strategies can supplement this method. Non-criterion-based strategies, such as inquiring about individuals' *expressed interests*, or directly stated self-reported interests or occupational goals, can inform interest assessment practice. Numerous structured and interactive approaches, such as versions of Tyler's (1961) vocational card sort intervention, focus on grouping cards indicating various occupations and academic majors into categories reflecting degree of interest (Hartung, 1999). Prominent examples of card sorts include the Missouri Occupational Card Sort (Krieshok, Hansen, & Johnston, 1989) and the Occupational Interest Card Sort (Knowdell, 1993).

Interest Inventory Scale Construction

Scales on interest inventories are constructed using one or more of three basic strategies: (a) the rational method, (b) the statistical method, and (c) the empirical method of contrast groups. To measure respondents' interests in broad, homogenous interest domains such as Holland's six types, a combination of rational and statistical methods usually is used. The rational method begins with a clearly specified definition of the domain, which guides the process of writing items that tap into the critical features of that construct. Once these items are in place, inventory developers use the statistical method to test the quality of the scale using data analytic strategies, such as factor analysis or cluster analysis, designed to identify underlying dimensions in an item pool. Assuming well-designed data collection procedures are used, the results of such statistical techniques can reveal whether the rationally derived items behave in the manner claimed of them, such as whether their scores assess homogenous domains where specified and whether ratings on all items contribute meaningfully to the total scores produced by each scale. This stage in scale construction often leads to the deletion of statistically unhelpful items from scales. Inventory developers also can assess whether scale scores correlate strongly with scores on other measures of the same or similar variables (convergent validity) and weakly with scores on dissimilar variables (discriminant validity).

The third strategy for scale construction, the empirical method of contrast groups, was developed and refined by E. K. Strong Jr., who was interested in successfully predicting satisfied membership in particular occupations. To accomplish this goal, he identified members of specific occupations and screened them to ensure that they were experienced in their fields and reported satisfaction with their jobs. Next, he identified an "in-general" criterion group consisting of people representing a wide range of occupations. Strong then administered a questionnaire in which respondents indicated liking, disliking, or indifference to a wide range of activities and occupations, which served as the items. To construct a particular occupational scale, Strong examined the percentages of like, indifferent, and dislike responses for men in the in-general group versus those in, say, the group of male lawyers. On some items there were clear differences in responses between the in-general and occupational group; these were the items selected for that occupation's scale. Once the scale was normed on the occupational sample, the scores produced by the scale for an individual respondent represent that person's degree of similarity (in terms of interests) to the happily employed women or men in that occupation. This type of scale is evaluated according to how well its scores successfully predict (concurrently and in the future) group membership, irrespective of the actual content of the items.

Popular Interest Inventories

Four of the most frequently used interest inventories are the SII (Donnay et al., 2005), the CISS (Campbell et al., 1992), the Self-Directed Search (Holland, 1985), and the KCS (Zytowski, 2009). The following sections describe each of these interest inventories in turn, illustrating their use (as well as the convergence of their scores) with a case example, a faux client named Soledad. In addition, numerous well-validated measures use similar measurement strategies, whereas others offer unique approaches, including content and norms for diverse groups. Space constraints preclude a thorough description of other prominent inventories currently available, including the Career Assessment Inventory (Johansson, 2002), the ACT Interest Inventory (Swaney, 1995), the Harrington O'Shea Career Decision Making System—Revised (O'Shea & Feller, 2009), the Jackson Vocational Interest Survey (Jackson, 1977), and the CAPA Interest Inventory (Borgen & Betz, 2008). Excellent resources are available for additional details examining the use of these and other measures (e.g., Hood & Johnson, 2007; Osborn & Zunker, 2006; Prince & Heiser, 2000; Whitfield et al., 2009).

The Case of Soledad

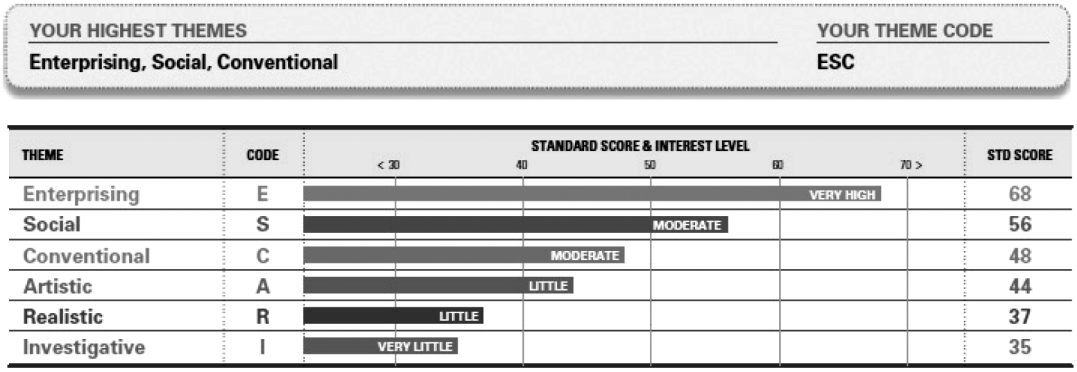
Soledad is a 45-year-old Argentinian American married mother of two with a bachelor's degree in business administration. She worked at the headquarters of a Fortune 500 tax and financial services firm for 12 years before being laid off because of changing

trends in service delivery driven by web-based services, fewer customers, and diminished revenues due to an economic recession. Soledad performed internal management consulting tasks related to logistics, purchasing, deployment, and distribution of products for business sites throughout the United States. Ongoing setbacks to this industry directly lessened the need for procuring materials for field offices. In an attempt to compete in this challenging business environment, Soledad's former employer restructured operations, which resulted in the elimination of hundreds of positions, ultimately flooding the local market with job seekers.

Soledad is in the throes of an unplanned midcareer transition stemming from broader market and technological forces. Support from outplacement services and severance pay afforded her an opportunity to explore other interests more related to her passions. She reported a complex mix of relief over leaving a struggling corporation and fear about an uncertain future. She noted that a thorough review of her interests, abilities, values, personality, and past accomplishments will help her identify a new career that will allow her to work with others in a more exciting field that better fits her needs for variety, prestige, and challenge, and that better satisfies her desire to learn marketable business skills. She reported longstanding interests in business, and leisure pursuits that include involvement in a book club, travel (especially tasks related to planning vacations), and jogging. As part of her work with a career counselor, Soledad completed a series of career assessments, including the following interest inventories: SII, CISS, SDS, and KCS.

The SII. The 2004 revision of the SII (Donnay et al., 2005), its most recent, consists of 291 items requiring respondents to indicate their interest in a wide range of occupations, subject areas, activities, and types of people using a 5-point scale ranging from *Strongly Like* to *Strongly Dislike*. Scores are reported on the SII profile for four sets of scales: six General Occupational Themes (GOTs), 30 Basic Interest Scales (BISs), 122 Occupational Scales (OSs), and five Personal Style Scales (PSSs). Administrative indices (e.g., an item response summary and a validity scale that screens for atypical response patterns) also are reported. The SII can be administered online or on paper, and a range of score reports are available from the publisher, CPP, Inc. (http://www.cpp.com), including high school and college editions, an expanded interpretive report, reports that combine SII scores with scores from the Myers–Briggs Type Indicator (MBTI; Myers, McCaulley, Quenk, & Hammer, 1998) or Skills Confidence Inventory (SCI; Betz, Borgen, & Harmon, 2005), and reports for use in organizations.

The GOTs on the Strong measure provide *T* scores ($M = 50$, $SD = 10$), representing the level of interests for the respondent relative to the SII's normative sample of employed women and men, for each of Holland's six types. Scores are displayed using a rank-ordered bar chart and numerical scores normed on the combined-sex general reference sample; verbal interpretive content (ranging from "very low" to "very high" based on the reference sample corresponding to the respondent's own sex) also is presented (see Figure 19.1). The first letters of Holland types among the highest three GOT scores are used to assign a Theme Code for the respondent, which can assist in interpreting patterns of results from other SII scores. The 30 BISs are homogenous scales that assess more specific interest domains (e.g., Performing Arts, Social Sciences, Research, Protective Services, Entrepreneurship, Finance & Investing). BIS scores also are represented with combined-sex-normed *T* scores depicted numerically and in bar charts and with verbal descriptors (abbreviated, ranging from VL [very low] to VH [very high]) normed on the relevant women-in-general or men-in-general sample (see Figure 19.2). Scores on the GOTs and BISs are supported by internal consistency reliabilities from .90 to .95 for the GOTs and .80 to .92 for the BISs, respectively; test–retest coefficients over intervals as long as 23 months range from .81 to .92 (GOTs) and from .74 to .93 (BISs; Donnay et al., 2005). GOT scores are intercorrelated in the pattern predicted by Holland's theory and correlate strongly with scores on other measures of like constructs. BIS scores correlate between .80 and .98 with the 1994 SII revision's BIS scores, and members of particular occupations tend to score high on BISs that are relevant for their job

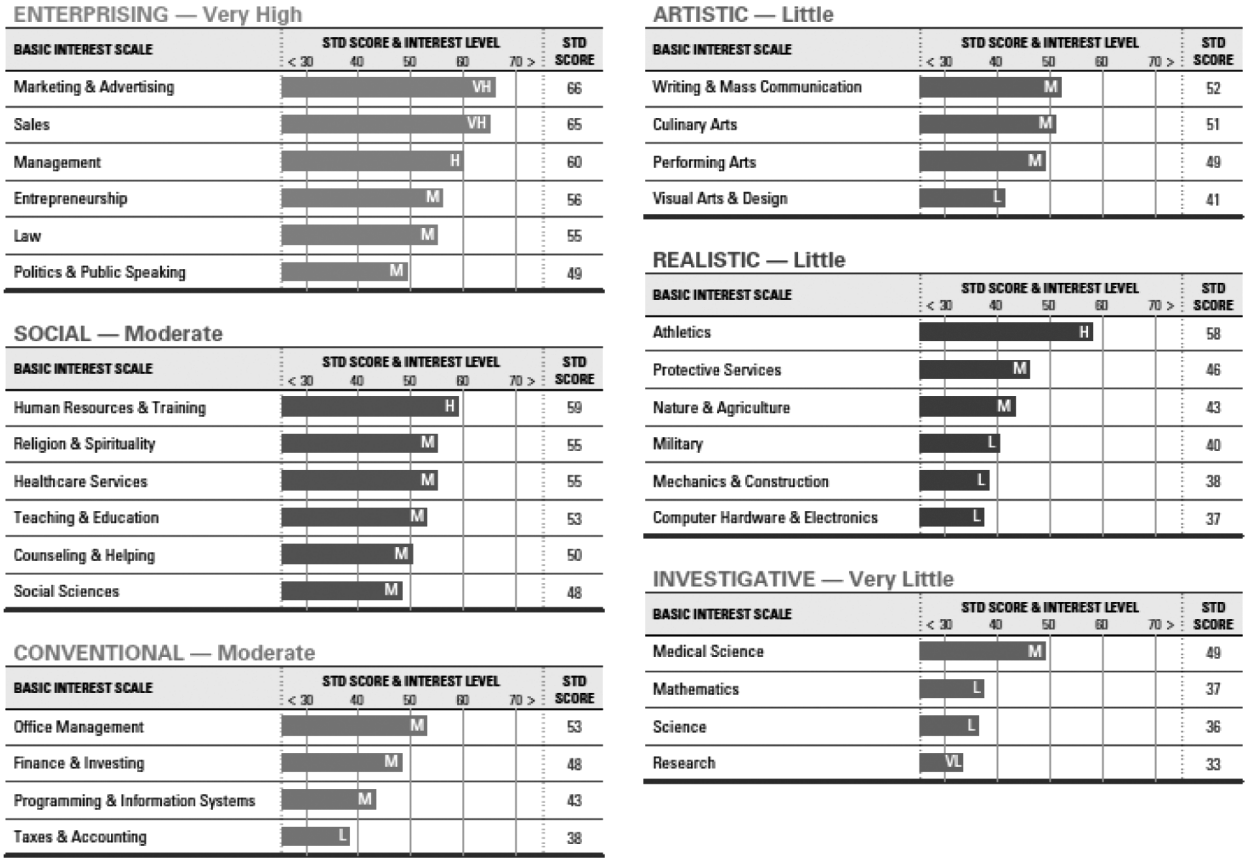


YOUR HIGHEST THEMES: Enterprising, Social, Conventional

YOUR THEME CODE: ESC

THEME	CODE	STANDARD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Enterprising	E	VERY HIGH	68
Social	S	MODERATE	56
Conventional	C	MODERATE	48
Artistic	A	LITTLE	44
Realistic	R	LITTLE	37
Investigative	I	VERY LITTLE	35

FIGURE 19.1. Soledad's General Occupational Theme scores on the SII profile. Modified and reproduced by special permission of the Publisher, CPP, Inc., Mountain View, CA 94043 from the Strong Interest Inventory Profile by CPP, Inc. Copyright 2004 by CPP, Inc. All rights reserved. Further reproduction is prohibited without the Publisher's written consent.



ENTERPRISING — Very High

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Marketing & Advertising	VH	66
Sales	VH	65
Management	H	60
Entrepreneurship	M	56
Law	M	55
Politics & Public Speaking	M	49

SOCIAL — Moderate

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Human Resources & Training	H	59
Religion & Spirituality	M	55
Healthcare Services	M	55
Teaching & Education	M	53
Counseling & Helping	M	50
Social Sciences	M	48

CONVENTIONAL — Moderate

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Office Management	M	53
Finance & Investing	M	48
Programming & Information Systems	M	43
Taxes & Accounting	L	38

ARTISTIC — Little

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Writing & Mass Communication	M	52
Culinary Arts	M	51
Performing Arts	M	49
Visual Arts & Design	L	41

REALISTIC — Little

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Athletics	H	58
Protective Services	M	46
Nature & Agriculture	M	43
Military	L	40
Mechanics & Construction	L	38
Computer Hardware & Electronics	L	37

INVESTIGATIVE — Very Little

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30 · 40 · 50 · 60 · 70 >)	STD SCORE
Medical Science	M	49
Mathematics	L	37
Science	L	36
Research	VL	33

FIGURE 19.2. Soledad's Basic Interest Scale scores on the SII profile. Modified and reproduced by special permission of the Publisher, CPP, Inc., Mountain View, CA 94043 from the Strong Interest Inventory Profile by CPP, Inc. Copyright 2004 by CPP, Inc. All rights reserved. Further reproduction is prohibited without the Publisher's written consent.

tasks, providing evidence of concurrent validity. Similarly, OS scores correlate in predicted directions with BIS scores, providing construct validity evidence (Donnay et al., 2005).

Soledad's GOT scores reveal a very strong interest in the Enterprising theme and a Theme Code of ESC (see Figure 19.1). This result suggests that Soledad was a good fit in terms of interests with her most

recent position, which likely satisfied many of her interests in business-related activity (e.g., sales, management, purchasing). Her BISs indicated very high interests in marketing and advertising and sales and high interests in management and human resources and training, but low interests in taxes and accounting, suggesting that it may have been the business-related tasks themselves that drew her to her previous job rather than the industry of tax and financial services. This finding suggests that when transitioning to a new position, the opportunity to engage in marketing, sales, management, and training activities may supersede the need that many laid-off individuals perceive to find another position in their same industry. (Alternatively, the experience of being laid off from a tax and financial services company may have influenced her interest ratings downward.) Soledad's Athletics score also was high, which may reflect her interest in jogging. Exploring whether Soledad feels her interest in athletics should be satisfied in her next job, rather than in her leisure time, may be useful; for many clients, a broader goal is to have their most salient interests satisfied within life as a whole, not necessarily solely on the job.

A total of 122 occupations are represented in the OSs of the SII (one OS for women and one for men for each occupation). The OSs were constructed using the empirical method of contrast groups (described earlier), with each OS normed on its own occupational criterion sample; this process yields *T* scores that indicate the degree to which the respondent's pattern of likes and dislikes corresponds to that of female or male workers representing that particular occupation. On the SII profile, the OSs are presented across three pages in bar chart format, with each OS listed under the Holland type represented in the first letter of the Theme Code derived from that occupation's mean GOT scores. Only the OSs normed on the respondent's own sex are depicted on the profile. Median test–retest reliability for OS scores, over intervals of up to 23 months, is .86; internal consistency reliabilities are not relevant because items were selected to differentiate between groups rather than to measure homogenous domains. The critical validity evidence for OS scores comes from predictive validity studies, which ask whether scores on OSs successfully predict satisfied membership in particular occupations. When pooled across earlier revisions of the instrument, evidence suggests that for intervals ranging from 3.5 to 18 years, approximately 65% of people had earlier scored at least 40 or higher on the OS corresponding to their current occupation in which they are happily employed. Soledad's OS scores under the Enterprising Theme (see Figure 19.3) suggest substantial similarity in terms of her pattern of interests with happily employed women who are retail sales managers, operational managers, restaurant managers, buyers, and so forth. Remarkably, most of the Enterprising OS scores for Soledad are higher than mid-range, suggesting a potentially wide range of Enterprise-themed occupations that she may find satisfying. As the profile indicates in the box to the right of her scores, Soledad also may benefit from examining the broader range of occupations presented in the O*NET.

The PSSs measure preferences for, or comfort with, particular aspects of work environments. They consist of Work Style (preferences or working alone vs. with people); Learning Environment (hands-on learning vs. traditional academic environment); Leadership Style (leading by example vs. directly taking charge of others); Risk Taking (playing it safe vs. enjoying risks); and Team Orientation (working independently vs. as part of a team on a shared project). These scales provide *T* scores normed on the combined-sex general reference sample and are supported by strong evidence of reliability (internal consistencies from .82 to .87, test–retest over 2- to 23-month intervals of .70–.91) and validity (based on scale intercorrelations and occupational and college major differences on relevant PSSs). Soledad's PSS scores suggested preferences for working with people (Work Style score of 69), hands-on learning (Learning Environment score of 42), taking charge of others (Leadership Style score of 57), and working on teams (Team Orientation score of 60); her Risk Taking score was a 49, suggesting she may not have a clear preference for taking risks, or that her risk tolerance may vary across situations.

The CISS. The CISS was introduced in the early 1990s by David Campbell, who was lead developer of the Strong measure during the 1960s and 1970s.

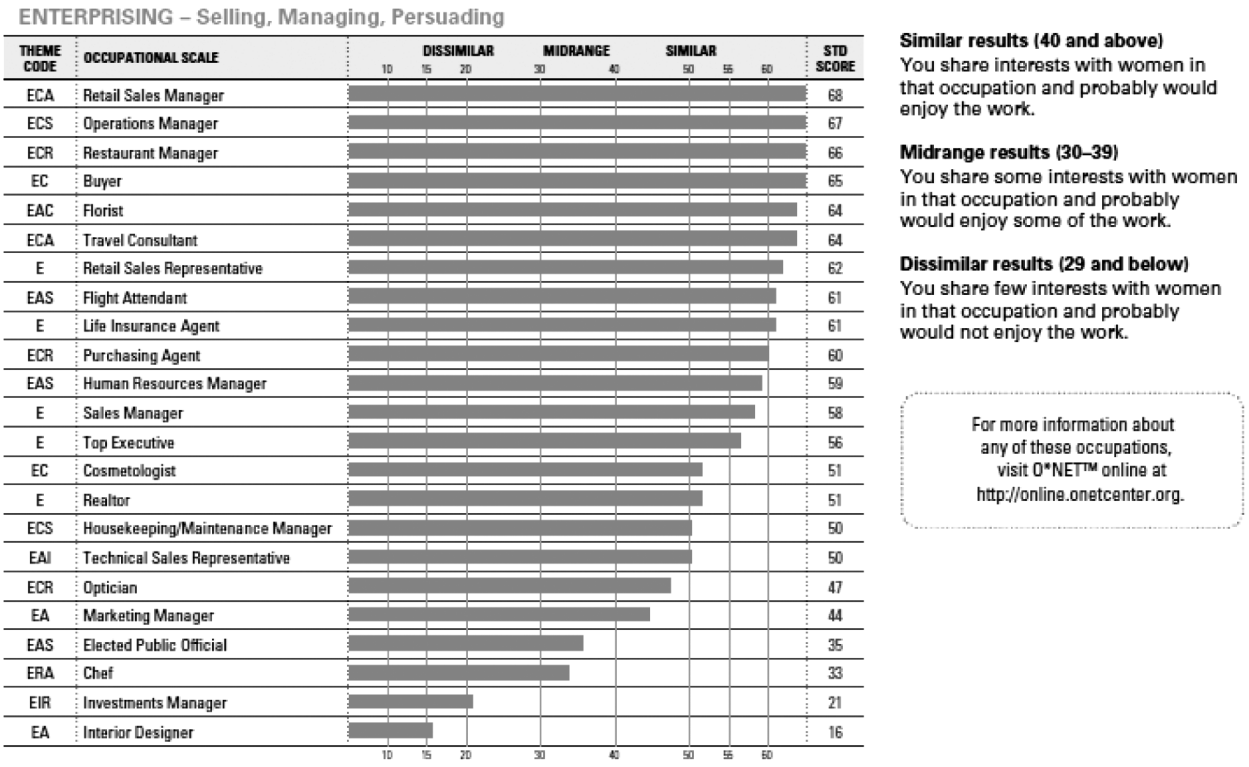


FIGURE 19.3. Soledad's Occupational Scale scores under the Enterprising Theme on the SII profile. Modified and reproduced by special permission of the Publisher, CPP, Inc., Mountain View, CA 94043 from the Strong Interest Inventory Profile by CPP, Inc. Copyright 2004 by CPP, Inc. All rights reserved. Further reproduction is prohibited without the Publisher's written consent.

The CISS is similar to the SII in that it provides interest scores using seven Orientation Scales, 29 Basic Scales, 60 Occupational Scales, and two Special Scales but adds the innovative feature of assessing self-estimated skills for the domains measured by the 98 interest scales. Respondents reply to 320 CISS items, all using a 6-point scale, ranging from *Strongly Like* to *Strongly Dislike* for the 200 interest items and *Expert, Good*, *Slightly Above Average*, *Slightly Below Average*, *Poor*, and *None* for the 120 skill items. The CISS can be administered on paper or online through the publisher, Pearson (http://psychcorp.pearsonassessments.com).

The Orientation Scales assess interests at the most global level on the CISS. The orientations are Influencing, Organizing, Helping, Creating, Analyzing, Producing, and Adventuring; these correspond to Holland's six types, with Holland's Realistic type represented by two CISS orientations, Producing and Adventuring. The 29 Basic Scales are similar to the BISs on the SII, although the each of the two sets of scales do present some interest domains not captured by the other (e.g., international activities and animal care on the CISS; data management and computer activities on the SII). These homogenous interest scales and their corresponding skills scales are presented on the 11-page CISS profile using *T* scores normed on a combined-sex general reference sample and are visually depicted along a *Very Low* to *Very High* continuum of interpretive comments. Elevations of the interest and skill scale for each orientation are compared in the scoring protocol and used to provide a recommendation of Pursue (if both interest and skill scores are high), Develop (high interest, low skill), Explore (low interest, high skill), or Avoid (low interest, low skill). For the Orientation Scales, median test–retest reliabilities are .87 for interest scales and .81 for skill scales, respectively, over a 90-day period; internal consistency reliabilities range from .82 to .93 (interest scales) and .76 to .89 (skill scales). Validity evidence for the Orientation Scales are derived

from their intercorrelations and relations with the SII GOT scores and, for the skills scales, a measure of self-estimated abilities (Hansen & Leuty, 2007; Sullivan & Hansen, 2004). Median alpha reliabilities for the Basic Scales are .86 (interest scales) and .79 (skill scales), and median 90-day test–retest reliabilities were .83 (interests) and .79 (skills). Evidence of validity for the Basic Scales includes large correlations with like scales on the SII and the utility of the scales for discriminating among people in different occupations in predictable ways (Campbell et al., 1992).

The CISS Occupational Scales were developed using the empirical method of contrast groups, similar to the OSs on the SII. Rather than providing separate scales for women and men, however, the CISS Occupational Scales were constructed using combined-sex criterion samples. To account for variability across sex, scores are calculated on the basis of the ratio of women to men in each occupation, in a manner that provides each sex with equal weighting in the conversion to standard scores. For each Occupational Scale, interest and skill *T* scores are normed on the general reference sample, although the profile also shows the range of scores for the middle 50% of satisfied members within the relevant occupation. The CISS Occupational Scale scores are supported by median 90-day test–retest reliabilities of .87 (interest scales) and .79 (skill scales), respectively (Campbell et al., 1992). Concurrent validity for the Occupational Scales are supported by evidence suggesting that 69% of women and 76% of men scored high on the interest scale corresponding to their college major; hit rates for the skill scales were 61% for both sexes, respectively (Hansen & Neuman, 1999). Finally, the CISS reports two special scales. The Academic Focus scale (interest and skill in academic pursuits, especially in the arts and sciences) was developed using the empirical method; its items are those that differentiated a diverse sample of highly educated people from less-educated individuals from the CISS occupational samples. The Extraversion scale (interest and skill in workplace activities requiring high levels of social interaction) consists of items that correlated strongly with observer ratings of extraversion.

As can be seen in Figure 19.4, Soledad's CISS profile suggests high interests in Helping, Influencing, and Organizing; her skills were lower than her interests in these domains—much lower in the case of the Helping orientation. Accordingly, the interpretive comments "Develop" and "Pursue" are found for several of the Basic Scales under these themes. This pattern was reversed for the Adventuring, Creating, Analyzing, and Producing orientations, in which her skill ratings exceeded her interests. However, because in most cases both skills and interests were low, "Avoid" was found for most of the scales associated with these orientations. Figure 19.5 shows Soledad's scores for the Occupational Scales under the Influencing orientation, many of which yield high scores and all but three of which are accompanied by the suggestion "Develop." Her counselor likely would explore with Soledad the fact that her skill estimates are lower than her interests in these domains; this may reflect a need for more experience, a pervasive sense of self-doubt in these areas (perhaps related to her layoff), or both. Soledad's special scale scores reflect high Extraversion (59 interest, 50 skill) and low Academic Focus (27 interest, 36 skill), suggesting that she is likely comfortable in social situations but prefers action-oriented and practical training to traditional academic environments. Interpretive comments on the profile note that "Business people, especially those in sales and marketing, tend to score low" on this scale.

The SDS. Holland's (1994b) Self-Directed Search—Form R is a 228-item instrument measuring the broad RIASEC types that can be administered, scored, and interpreted by high school, college and adult clients. The SDS was introduced in 1970 and subsequently revised in 1977, 1985, and 1994. It is available in several formats, including paper, computer, and online. Several additional forms are available for specialized populations, including individuals with limited reading skills (Form E; also available on audiotape), middle-school students (Form CE), and for professionals in organizational settings or other adults facing career transitions or seeking advancement (Form CP). The website of the publisher, PAR, Inc. (http://www4.parinc.com), reports that the SDS is available in

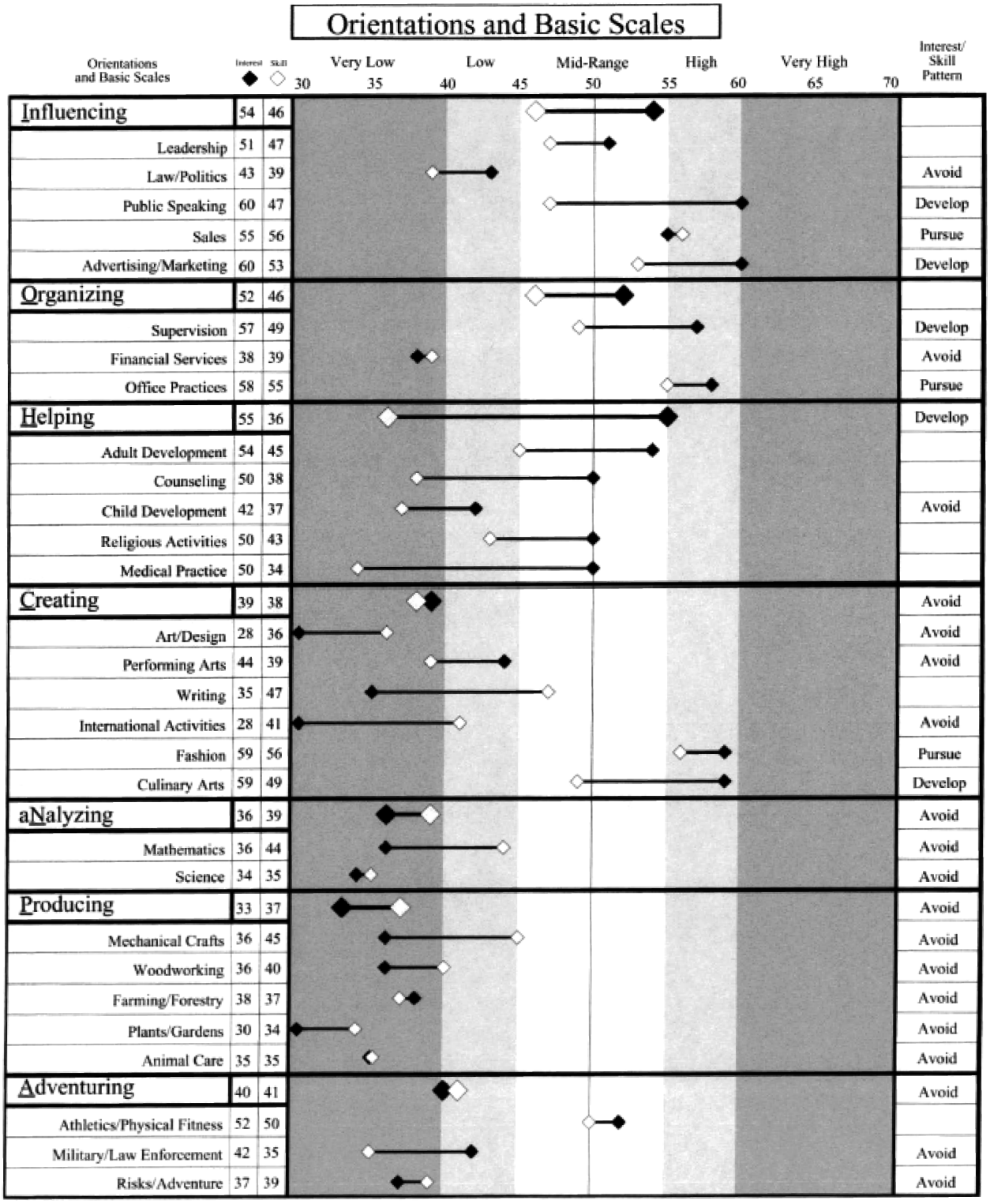


FIGURE 19.4. Soledad's CISS Orientations and Basic Scales. *Campbell Interest and Skill Survey (CISS)*. Copyright 1992 by David P. Campbell, PhD. Reproduced with permission of the publisher, NCS Pearson, Inc. All rights reserved. "*Campbell*" and "*CISS*" are trademarks of David P. Campbell, PhD.

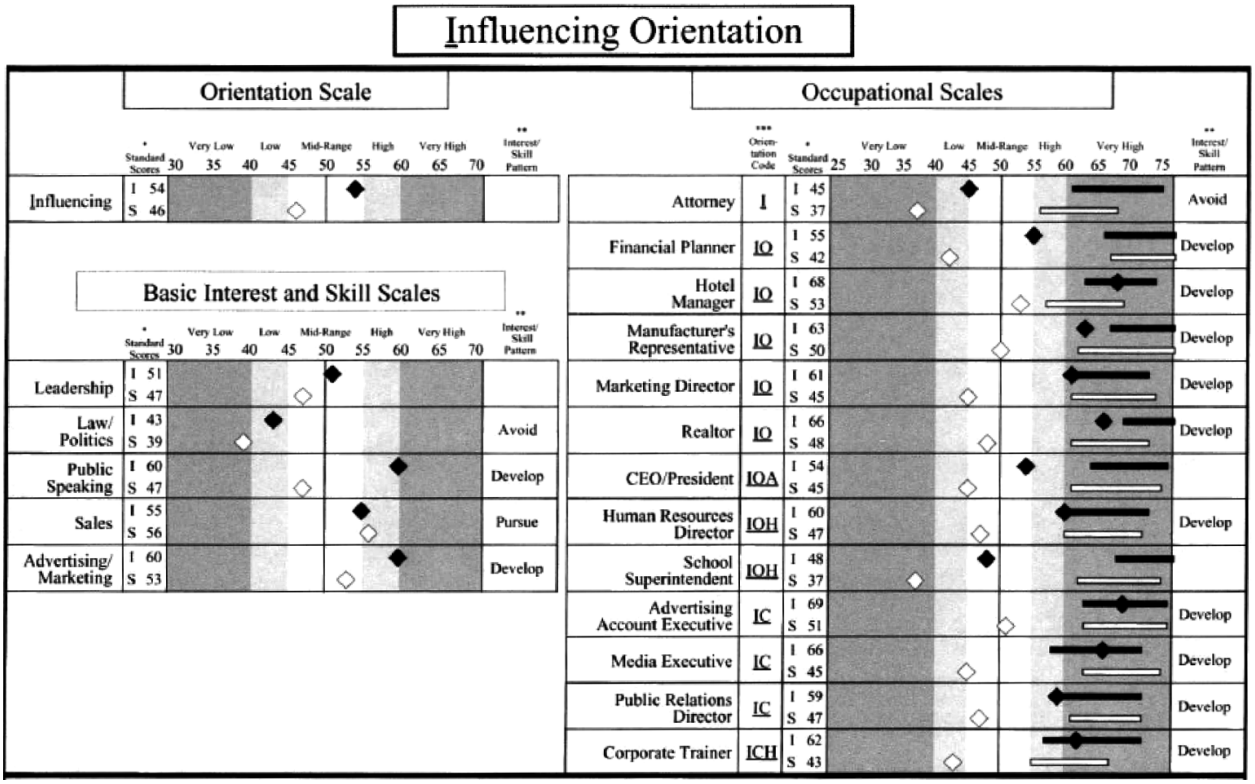


FIGURE 19.5. Soledad's CISS Orientation Scale score, Basic Scales, and Occupational Scales for the Influencing Orientation. *Campbell Interest and Skill Survey (CISS)*. Copyright 1992 by David P. Campbell, Ph.D. Reproduced with permission of the publisher, NCS Pearson, Inc. All rights reserved. "*Campbell*" and "*CISS*" are trademarks of David P. Campbell, PhD.

more than 25 languages and has been taken by more than 30 million individuals.

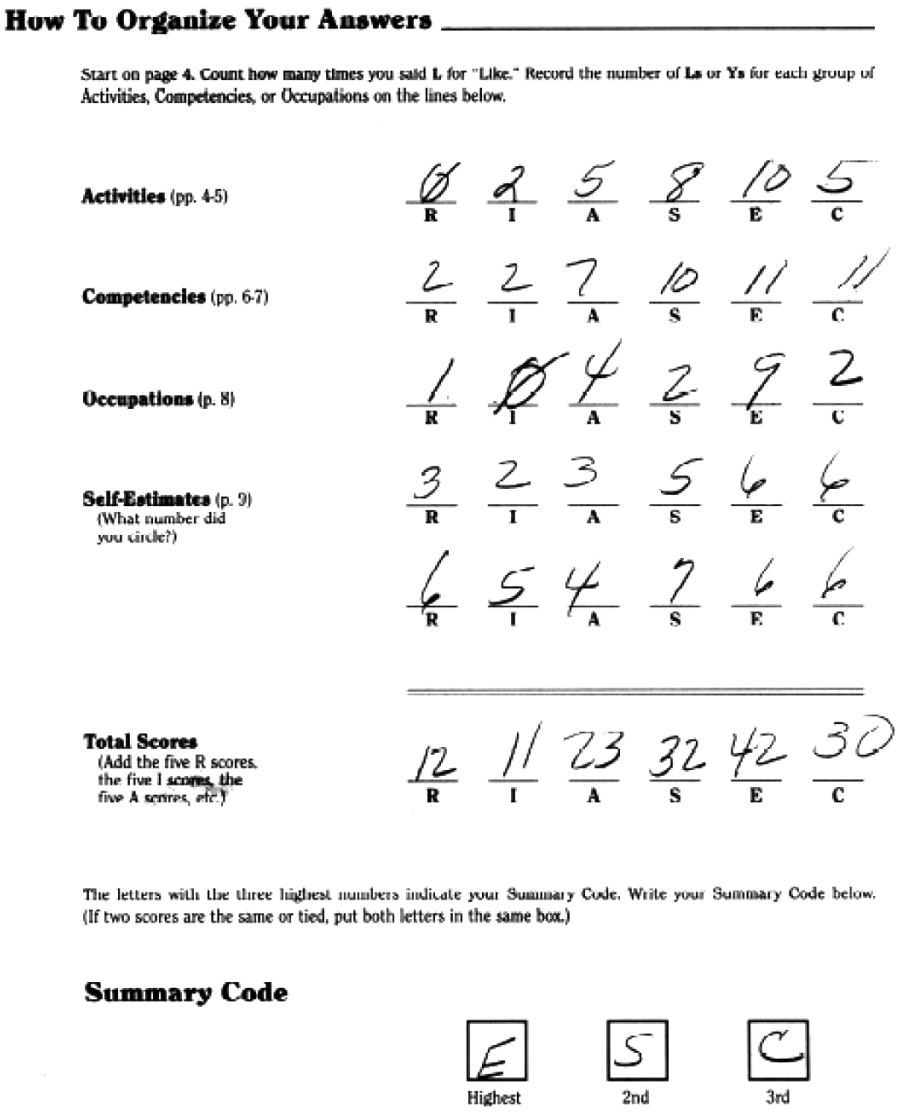
The SDS comprises five sections, including Occupational Daydreams and four scales assessing RIASEC types through Activities, Competencies, Occupations, and Self-Estimates, The Occupational Daydreams section is unique to the SDS and allows clients to list occupations they have considered throughout their lives. The response format differs slightly across each section: Like or Dislike for Activities; whether the respondent has performed well/competently (yes) or poorly (no) for Competencies; whether the occupation is interesting or appealing (yes) or uninteresting (no) for Occupations; and finally, respondents rate themselves "as you really think you are when compared with other persons your own age" on a scale ranging from 1 (*low*) to 7 (*high*), across two skills for each RIASEC domain for Self-Estimates. After completing the items, clients are instructed to sum items across all four sections for each Holland domain and list the top three letters as a Summary Code. Therefore, the resulting Holland Summary Code reflects a composite of interests in activities and occupations as well as self-estimates of competencies and abilities. This measurement approach is consistent with Holland's typological theory, in which scores reflect resemblance to each type composing a set of numerous qualities, and interests alone.

In the SDS Professional User's Guide, Holland, Powell, and Fritzsche (1997) reported strong Kuder–Richardson-20 internal consistency estimates, ranging from .72 to .92 for the subscales and .90 to .94 for the overall Summary Code scales. Test–retest stability correlations for the SDS scales are generally strong across 4 to 12 weeks, ranging from .76 to .89. Given that Holland type scores are operationally defined by the RIASEC scores on the SDS, numerous researchers have used this measure to evaluate the validity of Holland's theory.

Identifying a client's top three Holland codes is critical to interpreting SDS results. This step allows

clients to use detailed print and Internet-based resources to compare their Holland code results with related occupations, majors, and leisure activities. In addition to the SDS Interpretive Report, numerous guidebooks are available to assist clients in exploring options related to their results, including the *Dictionary of Holland Occupational Codes* (Gottfredson & Holland, 1996), *The Educational Opportunities Finder* (Rosen, Holmberg, & Holland, 1999), *The Occupations Finder* (Holland, 1994a), *The Alphabetized Occupations Finder* (Holland, 1997a), and *The Leisure Activities Finder* (Holmberg, Rosen, & Holland, 1999).

Soledad listed the following occupational daydreams with affiliated Holland codes: business owner (ESC), purchasing (ESR), pharmaceutical sales (EC), and project management in advertising and promotions (EAC). A summary of Soledad's results from each section is provided in Figure 19.6. Her overall Summary Code of ESC reflects a



How To Organize Your Answers

Start on page 4. Count how many times you said **L** for "Like." Record the number of **Ls** or **Ys** for each group of Activities, Competencies, or Occupations on the lines below.

	R	I	A	S	E	C
Activities (pp. 4-5)	0	2	5	8	10	5
Competencies (pp. 6-7)	2	2	7	10	11	11
Occupations (p. 8)	1	0	4	2	9	2
Self-Estimates (p. 9) (What number did you circle?)	3	2	3	5	6	6
	6	5	4	7	6	6
Total Scores (Add the five R scores, the five I scores, the five A scores, etc.)	12	11	23	32	42	30

The letters with the three highest numbers indicate your Summary Code. Write your Summary Code below. (If two scores are the same or tied, put both letters in the same box.)

Summary Code

Highest	2nd	3rd
E	S	C

FIGURE 19.6. Soledad's Self-Directed Search Summary Code results. Reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., 16204 North Florida Avenue, Lutz, FL 33549, from the Self-Directed Search Form R Assessment Booklet by John L. Holland, Ph.D., Copyright 1970, 1977, 1985, 1990, 1994. Further reproduction is prohibited without permission from PAR, Inc.

theoretically consistent profile, is congruent with most of her occupational daydreams, and can generalize to other occupations including human resource manager, community service manager, and loan counselor. Because of standard error of measurement among the six scales, the "Rule of 8" is often recommended (i.e., scores within eight points cannot be considered significantly different). Given that her top letter code of E (42) is 10 points higher than her second code, whereas S (32) and C (30) are within 8 points, she should also consider occupations with the ECS code. Finally, respondents are instructed to examine all five possible orders among these three codes; therefore, she may also consider the following codes: SEC, SCE, CSE, and CES. Her A code score also is within 8 points of the third highest letter, which supports the decision to review occupations with secondary A codes. The lowest scores—R (12) and I (11)—do not appear to match her interests very well. Her subscale RIASEC scores for the four sections reveal consistently low scores for R and I, except that she rated her abilities as above average for manual and math abilities. She endorsed almost all E type items across all section, mirroring her highly differentiated profile. As Soledad investigates occupations related to her Holland scores, she and her counselor would likely examine the job requirements and how they might relate to her present situation.

The KCS. Similar to E. K. Strong Jr., G. F. Kuder offered early innovations in interest measurement that continue to influence contemporary assessment approaches. First introduced in 1934 as the Preference Record, followed by the Kuder General Interest Survey (Kuder, 1975) and the KOIS (Kuder & Zytowski, 1991), the KCS (Zytowski, 2001a, 2009) represents the latest version. The KCS is part of a comprehensive online interactive career guidance program, known as the Kuder Career Planning System (KCPS), which also includes the Kuder Skills Assessment (Zytowski, Rottinghaus, & D'Achiardi, 2007), Super's Work Values Inventory—Revised (Zytowski, 2001b), and links enable clients to explore occupational and educational information related to assessment results. The KCS is considered a self-interpreting inventory, although the author encourages clients to use the support of counselors and teachers to maximize benefits. The KCPS offers three separate systems with developmentally appropriate activities and exploration features for elementary (Kuder Galaxy), middle school and high school (Kuder Navigator), and college/adult populations (Kuder Journey).

The KCS presents triads of items using a forced-choice method. Respondents rank order each list of three activities across 60 triads by indicating "Most" and "Least" preferred. Kuder (1977) believed that interest items should reflect activities and not occupational titles or school subjects emphasized in other inventories. In addition to College Major and Occupational Scales, the KOIS comprised 10 vocational interest estimates, which addressed the following basic interest domains: Artistic, Clerical, Computational, Literary, Mechanical, Musical, Outdoor, Persuasive, Scientific, and Social Service. These measures recently were updated as the 10 Activity Preference scales for the KCS as follows: Communications, Computational, Fine & Performing Arts, Managerial, Mechanical, Nature, Office Detail, Sales, Scientific, and Social Service. The Activity Preference measures typically are used as a hidden intermediate step for creating Person Match scores and scores on the 16 Career Clusters established by the U.S. Department of Education, Office of Vocational and Adult Education (http://www.careerclusters.org).

The set of 16 Career Cluster scales are developed using a unique criterion-keyed method that involves scoring weights for the 10 Activity Preference scales that separate large groups of adults employed in occupations reflecting the 16 Career Clusters. This multivariate approach yields weighted raw cluster scores that account for differences across the 10 scales distinguishing each identified group (e.g., Architecture & Construction) from the overall norm group of 8,791 individuals. Career Cluster scores are reported using percentile ranks "on the basis of their raw score distributions from the grand norm group" (Zytowski, 2009, p. 17).

The KCS introduced the unique Person Match method suggested by Kuder (1977, 1980). Kuder developed this approach because of a concern that individuals within any occupation are unique and,

therefore, mean scores representing a particular occupation are less meaningful given the diversity of activities and career patterns owing to the hetereogeneity among members. Instead of comparing respondents' patterns of responses to a group of individuals within a given occupation, as utilized in the occupational scales of the SII and KOIS, the KCS Person Match method involves "criterion individuals" taken from the larger normative sample. Person Match scores are derived by conducting rank-order profile comparisons for the 10 Activity Preference scales between respondents and each criterion pool member, yielding a Spearman rho correlation for each of the approximately 2,000 individuals in the Person Match pool. The top three Person Match scores within the respondent's top five career cluster domains are presented, with links to a detailed Person Match career story. These detailed stories provide biographical information and outline the individual's career development story through a series of questions examining typical work responsibilities, skills/attitudes necessary for success, pros and cons of the occupation, how they entered this occupation, future plans, and general advice (Zytowski, 2009; see Zytowski & D'Achiardi-Ressler, 2011, for an examination of the Person Match approach related to Markus and Nurius's, 1986, possible selves concept).

The technical manual for the KCS reports a series of studies examining the reliability and a detailed discussion on various forms of validity related to the Kuder inventories. The unique nature of the forced-choice response format and criterion-based measures presents challenges in reporting traditional psychometrics for the KCS. The KR-20 internal consistency reliability estimates of the 10 Activity Preference scales range from .64 (Nature) to .80 (Mechanical); 3-week test–retest stability coefficients range from .79 (Nature) to .92 (Art and Human Services). Correlations between the Kuder Career Clusters and Holland RIASEC scores from the SDS and SII reveal generally moderate relations for like-named scales.

Soledad's KCS profile (see Figures 19.7 and 19.8) highlights results organized by the 16 Career Clusters in addition to Person Match results. Zytowski (2009) encouraged clients to emphasize the rank ordering of career cluster scores, from most to least similar interests. Soledad's top three clusters—Business Management and Administration, Marketing, and Law, Public Safety, Corrections, and Security—all exceeded the 80th percentile, whereas her bottom three clusters—Agriculture, Food, and Natural Resources, Health Science, and Science technology, Engineering, and Mathematics—all fell below the 25th percentile. Soledad would likely benefit from exploring her top three clusters within the KCPS by clicking on links available within the profile. These clusters relate to the Holland Enterprising type, which appears to be rather pronounced across measures. Next, she would review her list of Person Match results and reflect on possible reasons for these results. It is interesting to note that her former position of "Purchasing Associate" was listed as a top Person Match. She would examine details for this sketch (and others) regarding activities, skills, educational level, and so forth. A practitioner can help her explore how the sketch answers relate to her current situation. She may also explore activities that help her consider her most and least preferred activities, work environments, and related occupations for her top three career cluster areas. The KCPS activities can help focus her career exploration process.

USING INTEREST INVENTORIES IN COUNSELING

Selecting an Inventory

Given the variety of uses and settings in which interest inventories are administered, practitioners must consider numerous factors when choosing an appropriate inventory. In addition to qualities (e.g., age, education, culture) of the test taker and reasons for taking the test (e.g., selecting a college major, job change, leisure counseling), the organizational setting also drives the decision-making process. For example, costs and administration setting must be taken into account by practitioners to achieve the most cost effective benefits for clients.

Since most inventories are routinely updated and revised, practitioners must pay attention to innovations and potential drawbacks of various versions.

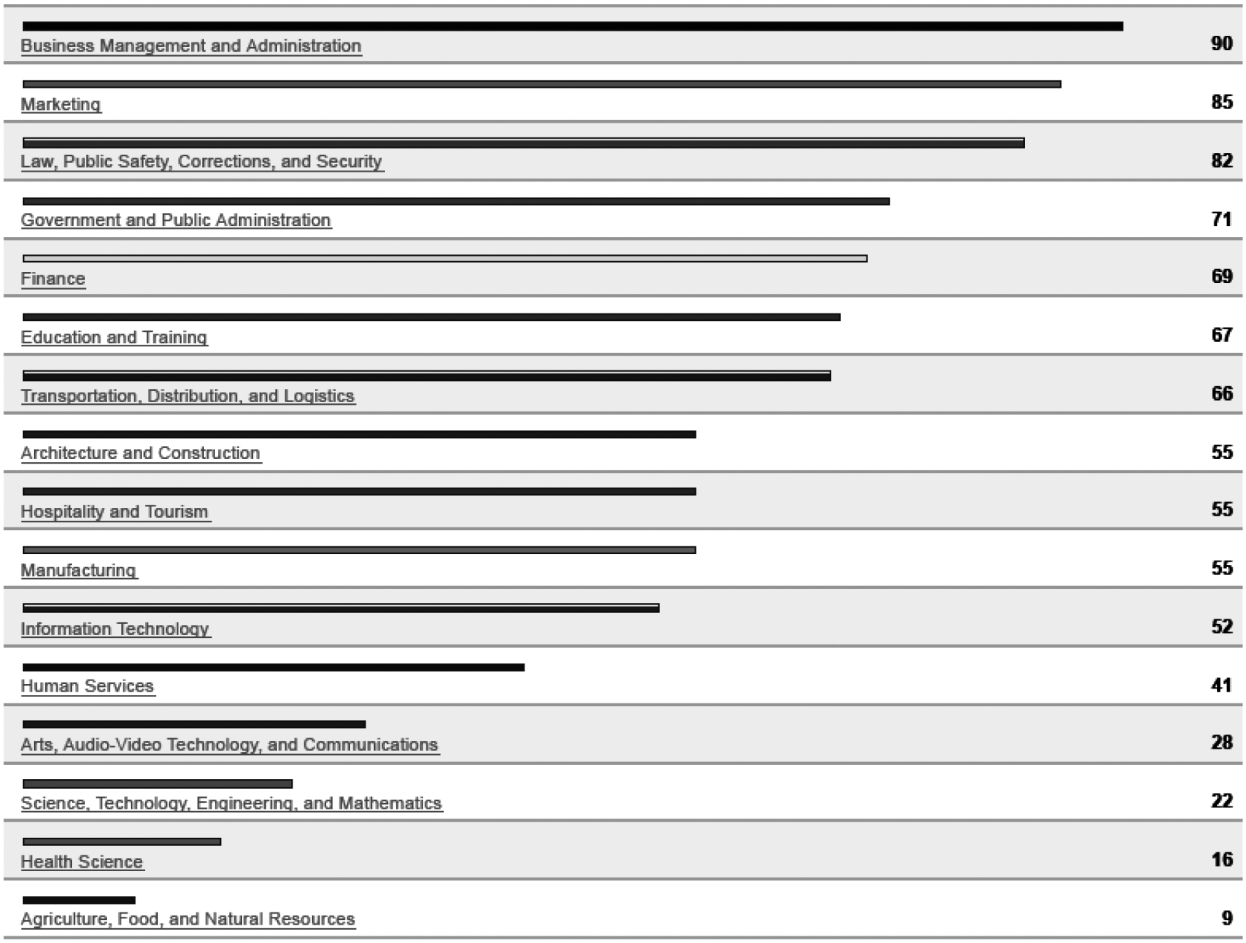


FIGURE 19.7. Soledad's Kuder Career Search with Person Match Career Cluster scores. Extracted and reproduced with permission of the publisher, Kuder, Inc. All rights reserved. Permission to reprint any portions of this extract must be sought from the publisher.

NAME	PERSON MATCH
Business Management and Administration	• Secretary #6 • Personnel Consultant • Pres., Large Manufacturing Company
Marketing	• Marketing Director • Sales Manager #1 • Purchasing Associate
Law, Public Safety, Corrections, and Security	• Attorney #7 • Attorney #8 • Corporate Lawyer
Government and Public Administration	• Urban Planner #1 • Economist • Transportation Planner
Finance	• Sales Executive, Insurance • Insurance Agent #2 • Cost Accountant #1

FIGURE 19.8. Soledad's Kuder Career Search with Person Match top three Person Match results for each of her top five Career Clusters. Extracted and reproduced with permission of the publisher, Kuder, Inc. All rights reserved. Permission to reprint any portions of this extract must be sought from the publisher.

Whitfield, Feller, and Wood's (2009) *A Counselor's Guide to Career Assessment Instruments* provides a good overview of commonly used inventories. Practitioners should consult technical manuals and user guides for information on psychometric properties (e.g., reliability, validity, norm groups) of the inventories and tips for effective interpretations. For example, the *Strong Interest Inventory Manual* (Donnay et al., 2005) includes a treasure of research on the Strong inventories throughout the years and interpretive strategies for a variety of profiles.

Practitioners also must consider supplemental resources, including affiliated assessments of other critical domains (e.g., skills, personality, values), interpretive guidebooks, and occupational information available for each inventory, particularly for online administrations. Computerized career-assisted guidance systems typically enable clients to connect their individualized results with online administrations of other inventories, which can be combined with detailed online occupational information resources such as the O*NET and the *Occupational Outlook Handbook* (Bureau of Labor Statistics, U.S. Department of Labor, 2010). Therefore, counselors should evaluate the needs of each client and plan to take advantage of these additional resources.

Administering an Interest Inventory

Clients may complete an inventory in an individual, small group, or classroom setting. Although many inventories can be self-administered (e.g., SDS, KCS), we encourage clients to explore the meaning of their results within the context of a formal professional relationship that includes additional interventions. Most interest measures are now available in hard copy and online formats that are connected to comprehensive computer-assisted career guidance systems. Professionals should provide a standardized introduction to help clients clarify specific reasons for testing, expectations, and potential benefits for the experience. Clients' ability to benefit from the overall intervention will depend on their understanding of these factors and awareness of additional print and online resources. Some clients will need additional assistance throughout testing because of limited reading comprehension or comfort using computers.

Preparing to Interpret an Interest Inventory

To prepare for the interpretation of an interest inventory with a client, counselors are urged to thoroughly review the profile ahead of time. First, the counselor should evaluate the validity of the instrument profile to ensure that the client approached the items in an open, honest, and consistent manner. In some circumstances, such as when reluctant clients are pushed to participate in counseling by a third party such as a parent or significant other, random responding may occur. Such validity concerns may be identified by a special scale, such as the Typicality Index on the SII or the Inconsistency Checks on the CISS, which assess consistency of responding to pairs of similar items. The total number of responses omitted—which, for many inventories, is noted on the profile report (or can be identified by scanning the item booklet)—also should be examined, as it can serve to identify fatigue or indifference. Next, the counselor should scan the breakdown of item responses to identify yea-saying or nay-saying response sets; many inventories (e.g., the SII and CISS) provide tables with response percentages to facilitate this.

Counselors can then turn to the scores to identify overall patterns of results, including special challenges for interpretation such as flat, depressed, or elevated profiles, or primary interest codes on opposite sides of Holland's hexagon. For inventories that provide scores representing different levels of specificity, the counselor should consider the relative consistency of score patterns across various types of scales, such as across occupational scales, basic interest scales, and scales assessing Holland's six types. For inventories that provide occupational scales constructed using the empirical method of contrast groups, counselors also may benefit from examining patterns of homogenous scale scores in light of high and low occupational scale scores. For example, surprisingly high occupational scale scores may be due to sharing dislikes more so than likes with an occupational criterion sample, a possibility that can be explored by examining the respondents'

patterns of scores on an inventory's homogenous scales. Similarly, high occupational scale scores should be examined in light of the clinical information already obtained from the client. Given that interest inventories unwittingly assess leisure interests as well as vocational interests, counselors can plan to explore which domains of life (e.g., work, leisure, or both) the client may choose to satisfy her or his interests. In this vein, the counselor can develop hypotheses regarding the client's interests profile, which can tested by raising questions with the client during the interpretation session.

Interpreting an Interest Inventory

Interest inventories can serve a variety of functions in the counseling process, such as helping clients efficiently identify their strongest areas of interest, uncovering previously foreclosed options and new directions to explore, and catalyzing discussion about which of several possible life domains (e.g., work, leisure, voluntarism) the client may choose to satisfy a particular set of interests. The counselor should have a thorough and accurate knowledge of the inventory in use, including how its scales were constructed and psychometric evidence, which provides the counselor with the ability to give the client a nuanced understanding of the instrument's scores. Such knowledge is necessary, but not sufficient for effective interpretation. Counseling skills such as building rapport, establishing a working alliance, providing an environment of safety and support, empathically attending to emotional responses to the information, eliciting the active engagement of the client, and connecting the interpretative data and client responses to the goals of counseling, also are critical (Hansen, 2005).

There is no one right way to interpret an interest inventory, but the following steps (see also Hansen, 2005) are representative of what counselors may cover in an interpretation session: (a) Review the goals for counseling and the purpose of incorporating the inventory into the counseling process in service of those goals. (b) Assess any reactions the client may have had about the experience of taking the inventory. (c) Orient the client to the nature of the information provided by the inventory—that is, it assesses interests and not other constructs (e.g., abilities, values, personality traits); it is not a crystal ball that tells a client what she or he "should" do but provides one source of information the client may use, along with other sources of information, to make informed choices about her or his career. (d) Build credibility for the inventory by briefly and very simply reviewing its history and summarizing research demonstrating the quality of the instrument. (e) Explain the theoretical framework (where relevant) for the inventory's scores (e.g., Holland's RIASEC model), and ask the client to predict her or his highest areas of interests using that framework. (f) Introduce the client to the score report, describe how scores are presented and how they may be interpreted (e.g., by noting how they are standardized and normed), and elicit the client's reaction using open-ended questions (i.e., "What is your reaction to this?"). (g) Identify and explore consistent patterns or interpretive challenges (i.e., "flat," elevated, depressed, or inconsistent profiles) that emerge across various scale scores, including their implications given the client's goals. (h) Work with the client to develop a strategy (e.g., searching the information provided by the O*NET) for using the information to identify good-fitting occupations and educational pathways that are not represented on the profile; (i) Provide a summary, or elicit a summary from the client, of the inventory's content and its relevance for the client's goals. (j) Coordinate a plan of action for the client to enact (i.e., homework) before the subsequent session.

SUMMARY

Vocational interests represent the core psychological space between people and many contexts of their lives. We have attempted to highlight advances in interest assessment throughout history, survey key research that elucidates validity of interest measures, and demonstrate how current approaches inform career counseling practice through the case of Soledad. These critical facets of individuality are influenced by numerous innate and contextual factors, and linked to career-related goals both theoretically and empirically. Leona Tyler (1978) considered interests as "possibility-processing structures" (p. 113) that support identity development and well-being.

Indeed, most career development theories emphasize the importance of interests and exploring the meaning of measured and expressed interests related to important life goals. Building on the strong foundation of this impressive body of literature, interest measures and assessment practices must continue to evolve to inform clients seeking meaning in an increasingly complex and uncertain world.

References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227–257. doi:10.1016/S0160-2896(96)90016-1

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219–245. doi:10.1037/0033-2909.121.2.219

Armstrong, P. I., Day, S. X., McVay, J. P., & Rounds, J. (2008). Holland's RIASEC model as an integrative framework for individual differences. *Journal of Counseling Psychology, 55*, 1–18. doi:10.1037/0022-0167.55.1.1

Armstrong, P. I., Hubert, L., & Rounds, J. (2003). Circular unidimensional scaling: A new look at group differences in interest structure. *Journal of Counseling Psychology*, *50*, 297–308. doi:10.1037/0022-0167.50.3.297

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology, 4*, 359–373. doi:10.1521/jscp.1986.4.3.359

Betsworth, D. G., Bouchard, T. J., Jr., Cooper, C. R., Grotevant, H. D., Hansen, J. C., Scarr, S., & Weinberg, R. A. (1994). Genetic and environmental influences on vocational interests assessed using biological and adoptive families and twins reared apart and together. *Journal of Vocational Behavior, 44*, 263–278. doi:10.1006/jvbe.1994.1018

Betsworth, D. G., & Fouad, N. A. (1997). Vocational interests: A look at the past 70 years and a glance at the future. *Career Development Quarterly, 46*, 23–47. doi:10.1002/j.2161-0045.1997.tb00689.x

Betz, N. E., Borgen, F. H., & Harmon, L. W. (2005). *Manual for the Skills Confidence Inventory* (Rev. ed.). Mountain View, CA: Consulting Psychologists Press.

Betz, N. E., & Rottinghaus, P. J. (2006). Current research on parallel measures of interests and confidence for basic dimensions of vocational activity. *Journal of Career Assessment, 14*, 56–76. doi:10.1177/1069072705281348

Borgen, F. H., & Betz, N. E. (2008). *Manual for the CAPA Interest Inventory*. Ames, IA: Career and Personality Assessments.

Borgen, F. H., & Lindley, L. D. (2003). Individuality and optimal human functioning: Interests, self-efficacy, and personality. In W. B. Walsh (Ed.), *Counseling psychology and optimal human functioning* (pp. 55–91). Hillsdale, NJ: Erlbaum.

Bubany, S., T., & Hansen, J. C. (2011). Birth cohort change in the vocational interests of female and male college students. *Journal of Vocational Behavior, 78*, 59–67. doi:10.1016/j.jvb.2010.08.002

Bureau of Labor Statistics, U.S. Department of Labor. (2010). *Occupational outlook handbook* (2010–2011 ed., Bulletin 2800). Washington, DC: U.S. Government Printing Office.

Campbell, D. P. (1966). The stability of vocational interests within occupations over long time spans. *Personnel and Guidance Journal, 44*, 1012–1019. doi:10.1002/j.2164-4918.1966.tb03825.x

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank*. Stanford, CA: Stanford University Press.

Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell Interest and Skill Survey: CISS*. Minneapolis, MN: National Computer Systems.

Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of Counseling Psychology, 39*, 7–19. doi:10.1037/0022-0167.39.1.7

Dawis, R. V. (2001). Toward a psychology of values. *The Counseling Psychologist, 29*, 458–465.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment*. Minneapolis: University of Minnesota Press.

Day, S. X., & Rounds, J. (1997). "A little more than kin, and less than kind": Basic interests in vocational research and career counseling. *Career Development Quarterly, 45*, 207–220. doi:10.1002/j.2161-0045.1997.tb00465.x

Day, S. X., & Rounds, J. (1998). Universality of vocational interest structure among racial and ethnic minorities. *American Psychologist, 53*, 728–736. doi:10.1037/0003-066X.53.7.728

Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 43–70). Hillsdale, NJ: Erlbaum.

Dik, B. J., & Hansen, J. C. (2008). Following passionate interests to well-being. *Journal of Career Assessment, 16*, 86–100. doi:10.1177/1069072707305773

Donnay, D. A. C. (1997). E. K. Strong's legacy and beyond: 70 years of the Strong Interest Inventory. *Career Development Quarterly, 46*, 2–22. doi:10.1002/j.2161-0045.1997.tb00688.x

Donnay, D. A. C., Morris, M. A., Shaubhut, N. A., & Thompson, R. C. (2005). *Strong Interest Inventory*

manual: Research, development, and strategies for interpretation. Palo Alto, CA: Consulting Psychologists Press.

Fouad, N. A., Harmon, L. W., & Borgen, F. H. (1997). The structure of interests in employed male and female members of U.S. racial/ethnic minority and nonminority groups. *Journal of Counseling Psychology, 44*, 339–345. doi:10.1037/0022-0167.44.4.339

Fouad, N. A., & Walker, C. M. (2005). Cultural influences on responses to items on the Strong Interest Inventory. *Journal of Vocational Behavior, 66*, 104–123. doi:10.1016/j.jvb.2003.12.001

Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology, 2*, 300–319. doi:10.1037/1089-2680.2.3.300

Freyd, M. (1923). *Occupational interests.* Chicago, IL: Stoelting.

Gottfredson, L. S. (2002). Gottfredson's theory of circumscription, compromise, and self-creation. In D. Brown, & Associates. (Eds.), *Career choice and development* (4th ed., pp. 85–148). San Francisco, CA: Jossey-Bass.

Gottfredson, L. S., & Holland, J. L. (1996). *Dictionary of Holland occupational codes* (3rd ed.). Lutz, FL: Psychological Assessment Resources.

Hansen, J. C. (1988). Changing interests: Myth or reality? *Applied Psychology: An International Review, 37*, 137–150.

Hansen, J. C. (2005). Assessment of interests. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 281–304). Hoboken, NJ: Wiley.

Hansen, J. C., & Dik, B. J. (2004). Measures of career interests. In M. Herson & J. C. Thomas (Eds.), *Handbook of psychological assessment: Vol. 4. Industrial/organizational assessment* (pp. 166–191). New York, NY: Wiley.

Hansen, J. C., & Leuty, M. E. (2007). Evidence of validity of the Skill Scale scores of the Campbell Interest and Skill Survey. *Journal of Vocational Behavior, 71*, 23–44. doi:10.1016/j.jvb.2007.04.006

Hansen, J. C., & Neuman, J. L. (1999). Evidence of concurrent prediction of the Campbell Interest and Skill Survey (CISS) for college major selction. *Journal of Career Assessment, 7*, 239–247. doi:10.1177/106907279900700304

Hartung, P. J. (1999). Interest assessment using card sorts. In M. L. Savickas & A. R. Spokane (Eds.), *Vocational interests: Their meaning, measurement, and use in counseling* (pp. 235–252). Palo Alto, CA: Davies–Black.

Hogan, R. T. (1983). A socioanalytic theory of personality. In M. Page (Ed.), *Nebraska Symposium on Motivation: Vol. 30. Personality: Current theory and research* (pp. 55–89). Lincoln: University of Nebraska Press.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*, 35–45. doi:10.1037/h0040767

Holland, J. L. (1985). *Self-Directed Search professional manual.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1994a). *The occupations finder.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1994b). *Self-Directed Search, Form R.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1997a). *The alphabetized occupations finder.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1997b). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1997). *Self-Directed Search professional user's guide.* Odessa, FL: Psychological Assessment Resources.

Holmberg, K., Rosen, D., & Holland, J. L. (1999). *The leisure activities finder.* Odessa, FL: Psychological Assessment Resources.

Hood, A. B., & Johnson, R. W. (2007). *Assessment in counseling: A guide to the use of psychological assessment procedures* (4th ed.). Alexandria, VA: American Counseling Association.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey manual.* Port Huron, MI: Sigma Assessments Systems, Inc.

Johansson, C. (2003). *Career Assessment Inventory—Enhanced Version.* San Antonio, TX: Pearson.

Knowdell, R. L. (1993). *Manual for occupational interests card sort kit.* San Jose, CA: Career Research and Testing.

Krieshok, T. S., Hansen, R. N., & Johnston, J. A. (1989). *Missouri occupational card sort manual.* Columbia: University of Missouri Career Planning and Placement Center.

Kuder, G. F. (1939). *Kuder Preference Record—Form A.* Chicago, IL: University of Chicago Bookstore.

Kuder, G. F. (1975). *Manual: Kuder E General Interest Survey.* Chicago, IL: Science Research Associates.

Kuder, G. F. (1977). *Activity interests and occupational choice.* Chicago, IL: Science Research Associates.

Kuder, G. F. (1980). Person matching. *Educational and Psychological Measurement, 40*, 1–8. doi:10.1177/001316448004000101

Kuder, G. F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey—Form DD general manual.* Monterey, CA: California Testing Bureau.

Langsdorf, P., Izard, C. E., Rayias, M., & Hembree, E. A. (1983). Interest expression, visual fixation, and heart rate changes in 2- to 8-month-old infants. *Developmental Psychology, 19*, 375–386. doi:10.1037/0012-1649.19.3.375

Larson, L., Rottinghaus, P., & Borgen, F. H. (2002). Meta-analysis of Big Six interests and Big Five personality factors. *Journal of Vocational Behavior, 61*, 217–239. doi:10.1006/jvbe.2001.1854

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79–122. doi:10.1006/jvbe.1994.1027

Low, D. K. S., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The stability of interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. *Psychological Bulletin*, *131*, 713–737. doi:10.1037/0033-2909.131.5.713

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." In S. T. Fiske (Ed.), *Annual review of psychology* (Vol. 51, pp. 405–444). Palo Alto, CA: Annual Reviews.

Markus, H., & Nurius, P. (1986). Possible selves. *American Psychologist, 41*, 954–969. doi:10.1037/0003-066X.41.9.954

Mitchell, L. K., & Krumboltz, J. D. (1996). Krumboltz's learning theory of career choice and counseling. In D. Brown & L. Brooks (Eds.), *Career choice and development* (3rd ed., pp. 233–280). San Francisco, CA: Jossey-Bass.

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *Manual for the Myers–Briggs Type Indicator* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Nauta, M. M., Kahn, J. H., Angell, J. W., & Cantarelli, E. A. (2002). Identifying the antecedent in the relation between career interests and self-efficacy: Is it one, the other, or both? *Journal of Counseling Psychology, 49*, 290–301. doi:10.1037/0022-0167.49.3.290

Osborn, D. S., & Zunker, V. G. (2006). *Using assessment results for career development* (7th ed.). Florence, KY: Cengage Learning.

O'Shea, A. J., & Feller, R. (2009). *Harrington-O'Shea Career Decision-Making System—Revised.* San Antonio: Pearson.

Parsons, F. (2005). *Choosing a vocation.* Broken Arrow, OK: National Career Development Association. (Original work published 1909)

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 21*, 259–287. doi:10.1016/0001-8791(82)90036-7

Prince, J. P., & Heiser, L. J. (2000). *Essentials of career interest assessment.* New York, NY: Wiley.

Reeve, J. (1993). The face of interest. *Motivation and Emotion, 17*, 353–375. doi:10.1007/BF00992325

Reeve, J., & Nix, G. (1997). Expressing intrinsic motivation through acts of exploration and facial displays of interest. *Motivation and Emotion, 21*, 237–250. doi:10.1023/A:1024470213500

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*, 3–25. doi:10.1037/0033-2909.126.1.3

Rosen, D., Holmberg, K, & Holland, J. L. (1999). *The educational opportunities finder.* Odessa, FL: Psychological Assessment Resources.

Rottinghaus, P. J., Betz, N. E., & Borgen, F. H. (2003). Validity of parallel measures of vocational interests and confidence. *Journal of Career Assessment, 11*, 355–378. doi:10.1177/1069072703255817

Rottinghaus, P. J., Coon, K., Gaffey, A., & Zytowski, D. G. (2007). Thirty-year stability and predictive validity of vocational interests. *Journal of Career Assessment, 15*, 5–22. doi:10.1177/1069072706294517

Rottinghaus, P. J., Lindley, L., Green, M. A., & Borgen, F. H. (2002). Educational Aspirations: The contribution of personality, self-efficacy, and interests. *Journal of Vocational Behavior, 61*, 1–19. doi:10.1006/jvbe.2001.1843

Rottinghaus, P. J., & Zytowski, D. G. (2006). Commonalities among adolescents' work values and interests. *Measurement and Evaluation in Counseling and Development, 38*, 211–221.

Rounds, J., & Tracey, T. J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology, 78*, 875–890. doi:10.1037/0021-9010.78.6.875

Rounds, J., & Tracey, T. J. (1996). Cross-cultural structural equivalence of RIASEC models and measures. *Journal of Counseling Psychology, 43*, 310–329. doi:10.1037/0022-0167.43.3.310

Rounds, J. B., & Smith, T. Hubert, L., Lewis, P., & Rivkin, D. (1998). *Development of Occupational Interest Profiles (OIPs) for the O*NET.* Raleigh: Southern Assessment Research and Development Center, Employment Security Commission of North Carolina.

Savickas, M. L. (2002). Career construction: A developmental theory of vocational behavior. In D. Brown & Associates (Eds.), *Career choice and development* (4th ed., pp. 149–205). San Francisco, CA: Jossey-Bass.

Savickas, M. L., & Spokane, A. R. (1999). *Vocational interests: Meaning, measurement, and counseling use.* Palo Alto, CA: Davies-Black.

Sheu, H., Lent, R. W., Brown, S. D., Miller, M. J., Hennessy, K. D., & Duffy, R. D. (2010). Testing the choice model of social cognitive career theory across Holland themes: A meta-analytic path analysis. *Journal of Vocational Behavior, 76*, 252–264. doi:10.1016/j.jvb.2009.10.015

Silvia, P. J. (2001). Interest and interests: The psychology of constructive capriciousness. *Review of General Psychology, 5*, 270–290. doi:10.1037/1089-2680.5.3.270

Silvia, P. J. (2006). *Exploring the psychology of interest*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195158557.001.0001

Spokane, A. R., & Decker, A. R. (1999). Expressed and measured interests. In M. L. Savickas & A. R. Spokane (Eds.), *Vocational interests: Their meaning, measurement, and use in counseling* (pp. 211–233). Palo Alto, CA: Davies-Black.

Strong, E. K., Jr. (1927). A vocational interest test. *Educational Record, 8*, 107–121.

Strong, E. K., Jr. (1943). *Vocational interests of men and women*. Palo Alto, CA: Stanford University Press.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*, 859–884. doi:10.1037/a0017364

Sullivan, B. A., & Hansen, J. C. (2004). Mapping associations between interests and personality: Toward a conceptual understanding of individual differences in vocational behavior. *Journal of Counseling Psychology, 51*, 287–298. doi:10.1037/0022-0167.51.3.287

Super, D. E. (1963). Self concepts in vocational development. In D. E. Super, R. Starishevsky, N. Matlin, & J. P. Jordaan (Eds.), *Career development: Self-concept theory* (pp. 1–16). New York, NY: College Entrance Examination Board.

Swaney, K. B. (1995). *Technical manual: Revised Unisex Edition of the ACT Interest Inventory (UNIACT)*. Iowa City, IA: ACT.

Swanson, J. L., & D'Achiardi, C. (2005). Beyond interests, needs/values, and abilities: Assessing other important career constructs over the life span. In S. Brown & R. Lent (Eds.), *Career development and counseling* (pp. 353–381). New York, NY: Wiley.

Tracey, T. J. G. (2002). Personal Globe Inventory: Measurement of the spherical model of interests and competence beliefs. *Journal of Vocational Behavior, 60*, 113–172. doi:10.1006/jvbe.2001.1817

Tracey, T. J. G., Watanabe, N., & Schneider, P. L. (1997). Structural invariance of vocational interests across Japanese and American cultures. *Journal of Counseling Psychology, 44*, 346–354. doi:10.1037/0022-0167.44.4.346

Tyler, L. (1961). Research explorations in the realm of choice. *Journal of Counseling Psychology, 8*, 195–201. doi:10.1037/h0041019

Tyler, L. (1978). *Individuality: Human possibilities and personal choice in the psychological development of men and women*. San Francisco, CA: Jossey-Bass.

Wallerstein, H. (1954). An electromyographic study of attentive listening. *Canadian Journal of Psychology/Revue canadienne de psychologie, 8*, 228–238. doi:10.1037/h0083613

Walsh, W. B. (1999). What we know and need to know: A few comments. In M. L. Savickas & A. R. Spokane (Eds.), *Vocational Interests: Meaning, measurement and counseling use* (pp. 371–382). Palo Alto, CA: Davies-Black.

Walsh, W. B., & Betz, N. E. (1995). *Tests and assessment* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Weick, K. E. (1964). Reduction of cognitive dissonance through task enhancement and effort expenditure. *Journal of Abnormal Psychology, 68*, 533–539. doi:10.1037/h0047151

Whitfield, E. A., Feller, R., & Wood, C. (2009). *A counselor's guide to career assessment instruments* (5th ed.). Broken Arrow, OK: National Career Development Association.

Zytowski, D. G. (2001a). Kuder Career Search with person-match: Career assessment for the 21st century. *Journal of Career Assessment, 9*, 229–241. doi:10.1177/106907270100900302

Zytowski, D. G. (2001b). *Super's Work Values Inventory, Revised: User's manual*. Adel, IA: National Career Assessment Services.

Zytowski, D. G. (2009). *Kuder Career Search With Person Match—Technical manual*. Retrieved from http://www.kuder.com/downloads/kcs-tech-manual.pdf

Zytowski, D. G., & D'Achiardi-Ressler, C. (2011). Person match as a source of possible selves. In P. J. Hartung & L. M. Subich (Eds.), *Developing self in work and career: Concepts, cases, and contexts* (pp. 109–121). Washington, DC: American Psychological Association. doi:10.1037/12348-007

Zytowski, D. G., Rottinghaus, P. J., & D'Achiardi, C. (2007). *The Kuder Skills Assessment user manual*. Adel, IA: Kuder.

CHAPTER 20

ASSESSMENT OF CAREER DEVELOPMENT AND MATURITY

Jane L. Swanson

The concept of career development emerged from the work of Donald E. Super (1953, 1980), in recognition of the ongoing nature of vocational decisions across a person's life span. Super's theory led to efforts to measure constructs unique to this approach, including vocational maturity (Super, 1955), career maturity (Crites, 1972) and more recently, career adaptability (Savickas, 1997; Super & Knasel, 1981). Additional constructs related to career development include career planning, career awareness, career aspirations, career exploration, career stages, career salience, career beliefs and thoughts, career decision-making styles and self-efficacy, and career indecision (Chartrand & Camp, 1991; Swanson & D'Achiardi, 2005). Taken together, these constructs have assumed a central role in theory, research and practice within vocational/career psychology, and efforts to adequately measure such constructs have occupied much attention for 60 years. Measuring such constructs has proved difficult, however, in part because of the inherently transitory nature of a construct expected to change with time.

This chapter presents the history and evolution of constructs related to career development and discusses specific measures. It also summarizes psychometric difficulties and potential new directions for assessment of career development constructs.

PURPOSE OF ASSESSING CAREER DEVELOPMENT

Generally speaking, psychological assessment serves several purposes. Measuring important characteristics of individuals aids in diagnosis and treatment, allows us to compare individuals to one another, provides a common language to communicate important information, and supplies objective and standardized information as a basis for decision making (Swanson, 2012). In the case of constructs related to career development, there are compelling theoretical and practical reasons for using well-developed and psychometrically sound measures.

From a theoretical perspective, well-designed measures of central constructs are necessary to advance theories of how individuals progress through their careers by allowing adequate testing of hypotheses derived from theory. From a practical standpoint, "determining clients' readiness to make educational or vocational choices is the *principal* assessment task in comprehensive career counseling" (Savickas, 2000, p. 427, italics added). A typical goal of career counseling is to assist clients in decisions related to career choice and implementation (Swanson & Fouad, 2010). Career counseling will likely be most effective with clients who are ready to make a decision, but it may not be useful at all if clients are not sufficiently ready.

Early in the emerging theory and research related to career maturity, Crites (1965) drew a distinction between career choice content and career choice process. Career choice content refers to the actual choice, or the *what*, or product, of career decision making; career choice process refers to *how* the decision is made (Crites, 1965; Savickas, 2005). This distinction becomes particularly important in considering the purpose of assessment. Assessment that

DOI: 10.1037/14048-020
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

focuses on the content of career development includes measures of interests, values, abilities, skills, and personality. All of these instruments are used for the client; in other words, the information gleaned from the assessment is shared directly with the client and becomes interwoven with what occurs in career counseling or educational interventions. This type of assessment is most amenable to the recommendations of Duckworth (1990) related to the client's involvement in all aspects of assessment, from initial selection of measures through interpretation of scores. Moreover, assessment of career choice content is frequently viewed as an intervention in and of itself within career counseling.

In contrast, the assessment of career choice process often serves a screening function to determine whether the client is ready to move forward with career decisions and to identify factors that may impede decision making. Assessment of career choice process also is frequently used as a criterion by which to judge the efficacy of interventions. Thus, measures of process may be more oriented toward use primarily by counselors than are measures of content, although some recent test developers have made materials more directly accessible to clients.

HISTORY AND EVOLUTION OF CAREER DEVELOPMENT CONSTRUCTS

The activities of vocational guidance and career counseling are often traced to the work of Frank Parsons (1909), in which he advocated for the systematic and rational analysis of an individual's strengths, knowledge of the requirements of different occupational paths, and the application of "true reasoning" to arrive at a match between the two domains. Parsons's work led to the development of inventories to measure relevant characteristics of individuals (e.g., interests, values, and abilities), as well as thorough descriptions of the range of occupations, culminating in the trait-and-factor approach to vocational guidance which predominated until the middle of the 20th century.

The history of vocational psychology is strongly tied to the emergence of the psychometric movement in the first half of the 20th century (Dawis, 1992), with measures of vocationally relevant constructs assuming a prominent role for researchers and practitioners. Development of measures of vocational interests, in particular, was prevalent in the middle third of the 20th century (see Chapter 19, this volume); measurement of values and needs also garnered attention (see Chapter 21, this volume). The focus of this effort was the accurate measurement of an individual's characteristics at a given, albeit static, point in time. Many of these characteristics were assumed to be relatively stable, at least sufficiently stable to use in making career-related decisions, and in general such assumptions have been supported (Hansen, 2005; Rounds & Armstrong, 2005). Results from these assessments were then used to assist individuals in making career choices based on the concurrent fit between their characteristics and those of the occupational world, and to predict future satisfaction and performance. Career decision making was viewed as a discrete event, a one-time occurrence that could be improved via the actuarial information provided by measures of interests, values, needs, and aptitudes.

In part fueled by work in fields other than psychology, the concept of vocational development began to emerge within the vocational guidance movement. Ginzberg, Ginsburg, Axelrad, and Herman (1951) described the nondiscrete nature of vocational choice, suggesting that individuals actually progressed through three stages (fantasy, tentative, and realistic) in making a decision. However, Donald Super was the primary impetus for the use of developmental concepts to describe vocational choice. In 1955, Super noted that the literature related to vocational guidance did not contain the term *vocational development*, instead focusing on *vocational choice*, implying a single event or point in time. In this seminal publication, Super discussed the advantages of borrowing from the discipline of developmental psychology, particularly the concept of individuals' vocational lives progressing through a series of stages, each characterized by unique tasks and challenges. Super (1953) posited five stages of vocational development that occurred across the life span: growth, exploration, establishment, maintenance, and decline; moreover, he introduced the concept of career pattern, or the sequence, frequency,

and duration of jobs over the life span. Furthermore, Super argued, the concept of vocational development "leads logically" to the concept of vocational maturity—the degree of vocational development that one has achieved on a continuum demarcated by the five stages outlined in his theory (Super, 1955, p. 153).

According to Super (1955), vocational maturity had a number of components: (a) orientation to vocational choice, (b) information and planning about one's preferred occupation, (c) consistency of vocational preference, (d) crystallization of traits, and (e) wisdom of vocational preferences. Super also proposed indices of vocational maturity, likening it to the measurement of intelligence as a ratio of mental age to chronological age. For example, he defined one vocational maturity quotient as the ratio of vocational maturity to chronological age, thus quantifying whether the vocational development of an individual was appropriate for his/her age, and the degree to which an individual deviated (above or below) from his/her age. As Super himself noted, he thus defined two important aspects of vocational maturity—the status of a person on a behavioral scale of development and his/her status relative to his/her age; so "vocational development is a continuum, and vocational maturity [is] a point on this continuum denoting degree of development attained" (Super, 1955, p. 154).

Thus began a 55-year quest to define the components of vocational maturity (later renamed *career maturity*) and then to measure the components with enough precision to be meaningful. Within a few years of Super's (1955) influential publication, other researchers began to write about conceptual and psychometric difficulties (Bartlett, 1971; Crites, 1961, 1974), which were echoed throughout the next several decades (Betz, 1988; Chartrand & Camp, 1991; Westbrook, 1983). Unfortunately, some of the thorniest psychometric issues remain unresolved, in part because of the difficulty of measuring developmental constructs, particularly those that are theorized to begin in adolescence and extend across the life span.

Attention to the developmental nature of career choice had a profound effect on the field of vocational psychology and, relatedly, to the assessment of constructs deemed important to theory, research, and practice (Phillips & Pazienza, 1988). Work by Super and Crites, as well as later researchers, highlighted the developmental nature of career activities throughout the life span, and led to interest in instruments that would capture the progression of career-related activities and tasks. Measurement of career maturity was at the forefront of this interest.

Super et al. reported results from the Career Pattern Study (1957; Super & Overstreet, 1960), including a further explication of the construct of vocational maturity to include 77 variables, distributed across six dimensions and 20 indices. In response, John Crites (1961) published an analysis of the conceptual underpinnings of vocational maturity, suggesting a need for a "reduction of the definitions linguistically as well as numerically before the initiation of further research" (p. 255), particularly prior to the development of a measure of vocational maturity. He argued that there were two independent and measurable constructs: degree of vocational development, an individual's progress in comparison with others in his or her life stage; and rate of vocational development, an individual's progress relative to others of the same age. An individual's level of vocational maturity thus should be gauged by both specific behaviors and completion of developmental tasks. Crites (1965) also reorganized the dimensions identified through Super's Career Pattern Study into a hierarchical model of vocational maturity, containing 18 dimensions in two content factors (wisdom of choice and consistency of choice) and two process factors (choice competencies and choice attitudes).

Following the work of Super and Crites, there was a flurry of research focusing on career maturity, including evaluation of the instruments that they developed (the Career Development Inventory [CDI] and the Career Maturity Inventory [CMI], respectively) as well as development of additional instruments. For example, Westbrook (1970, cited in Westbrook, 1983) developed the Cognitive Vocational Maturity Test to focus solely on the cognitive aspects of the construct, with six scales tapping knowledge of the world of work, and Gribbons and Lohnes (1968) developed a structured interview titled the Readiness for Career Planning. Several

measures were also developed specifically for use with adults, including the Adult Vocational Maturity Inventory (Sheppard, 1971), the Career Adjustment and Development Inventory (Crites, 1979, cited in Betz, 1988), and the Adult Career Concerns Inventory (ACCI; Super, Thompson, Lindeman, Myers, & Jordaan, 1985).

Continued research led to additional concerns about conceptual issues as well as psychometric concerns about available instruments. Westbrook (1983) raised a number of issues related to reliability and validity of extant measures, including lack of consensus about the number or structure of dimensions; substantial variation in content; lack of convergent and discriminant validity among measures of career maturity and other constructs; and lack of sufficient reliability evidence. Furthermore, he concluded that a stronger research base was necessary before measures of career maturity could be used for differential diagnosis in career counseling and that few studies using career maturity as a criterion measure reported significant results. Similarly, Betz (1988) concluded that there was a lack of agreement regarding both the construct and the consequences of career maturity and that measures evinced "relatively poor" reliability and validity (p. 117). The state of the literature at that time led Westbrook et al. (1980) to conclude that "the concept of career maturity is an endangered species" (p. 278).

In a later review of the measurement of career development constructs, Chartrand and Camp (1991) concluded that there were "still some vexing measurement problems that plague most career maturity measures" (p. 10); furthermore, they noted that "reliability tends to be problematic" and that there were "content discrepancies" across measures. They also argued that the greatest research need was attention to criterion-related validity. Despite such a pessimistic view, the construct of career maturity continued to occupy a central role in research and practice, and its recent evolution into career adaptability suggests a contemporary role.

From Career Maturity to Career Adaptability

In Super's initial conceptualization, the construct of career maturity was useful in explaining adolescent career development; however, it is less useful for understanding adult career development, given its emphasis on age-appropriate behavior (Savickas, 1997; Super & Knasel, 1981). In other words, whereas Super's theory focused on life-span development, the construct of career maturity did not. To address this issue, Super and Knasel (1981) introduced the construct of career adaptability, consisting of five elements: (a) planfulness, the ability to learn from experiences and anticipate the future; (b) exploration, the ability to ask questions and collect information and to interact with others; (c) information gathering, the ability to gather information about the world of work; (d) decision making, the ability to make choices; and (e) reality orientation, the ability to develop self-awareness, self-knowledge, and establish realistic options consistent with preferences (Super, 1983).

More recently, Savickas (1997) defined career adaptability as "the readiness to cope with the predictable tasks of preparing for and participating in the work role and with the unpredictable adjustments prompted by changes in work and working conditions" (p. 254). In his view, career adaptability encompasses three major components: "planful attitudes, self and environmental exploration, and informed decision making" (Savickas, 1997, p. 254). Furthermore, he described career adaptability as an integrative or bridging construct within "life-span, life-space theory" and proposed that it replace career maturity as the theory's central construct. Career adaptability thus allows the use of one construct across the life span. Recent research has suggested that the construct of career adaptability contributes to understanding adolescent and adult career development (cf. Creed, Fallon, & Hood, 2009; Hirschi, 2009; Koen, Klehe, Van Vianen, Zikic, & Nauta, 2010). However, as of yet, there is no established or accepted measure of career adaptability that is available for practitioners, although some show promise. Although researchers have determined ways to operationalize career adaptability, primarily by means of scales from measures such as the CMI and CDI, no standardized measure has yet been developed that would provide practitioners with normative information about clients. Additional attention to measuring career adaptability is necessary.

Additional Constructs Related to Career Development

The constructs of vocational/career maturity and career adaptability are at the heart of the developmental concept of growth over time, and, indeed, attention to measuring these constructs has advanced theory and practice related to career development. In addition to these constructs, however, are several other related constructs. These constructs all have a developmental flavor because they occur as individuals are faced with tasks throughout the life span, regardless of life stage (Chartrand & Camp, 1991; Swanson & D'Achiardi, 2005); furthermore, some of these constructs are subsumed in the current thinking about, and measurement of, the construct of career adaptability. Sampson, Peterson, Reardon, and Lenz (2000) classified all of these constructs under the rubric "readiness assessment."

Several constructs have been identified and studied related to the process of decision making. These constructs address how people make career decisions (decision-making style); the precursors that may influence or impede career choice (career indecision); and individuals' beliefs that they can successfully accomplish behaviors that will lead to designated outcomes (decision-making self-efficacy beliefs; Phillips & Jome, 2005; Swanson & D'Achiardi, 2005). Chartrand and Camp (1991) suggested that the study of career decision making was a "microanalysis of career development" (p. 10), by focusing on specific processes by which career decisions are made. In addition to these three constructs related specifically to aspects of decision making, two other constructs particularly relevant to career development are included herein; namely, (a) career salience and (b) career beliefs and thoughts.

Career decision-making styles. Different styles of decision making have been described by Jepsen and Dilley (1974), Harren (1979), and others. For example, Harren proposed three styles: rational (intentional and logical), intuitive (based on feelings and emotional responses), and dependent (made in accord with others' opinions). A recent approach (Gati, Landman, Davidovitch, Asulin-Peretz, & Gadassi, 2010) moves away from a typology to a decision-making profile in which an individual's behavior is classified on 11 separate dimensions.

Career indecision and other decision-making difficulties. Early research on career decision making dichotomized individuals as either decided or undecided; however, later research further explicated the construct of indecision and the factors that lead to difficulties in decision making (Slaney, 1988). An important advance was to isolate chronic or trait indecision, a "pervasive and enduring form of indecision that does not abate as information is acquired" (Kelly & Lee, 2002, p. 307) from other types of indecision. There seems to be general consensus that indecision is multidimensional; however, the specific nature and number of dimensions is still in question (Lonborg & Hackett, 2006). Another approach is Gati's (Gati, Krausz, & Osipow, 1996) hierarchical taxonomy of decision-making difficulties, in which he distinguished between difficulties that occur before the career decision-making process (lack of readiness) and those that occur during the process (lack of information, inconsistent information).

Career decision-making self-efficacy. A third decision-making construct receiving considerable attention is Taylor and Betz's (1983) application of Bandura's (1977) social learning theory. Career decision-making self-efficacy, defined by Taylor and Betz as an individual's confidence in his or her ability to complete career decision-making tasks effectively, has become a predominant outcome variable of studies examining career interventions.

Career salience. The construct of career salience was originally identified by Super, relevant to his "life-span, life-space" theory given the importance of life roles. Savickas (2002) described career salience as one of the major components of career adaptability, defined as the value and preference that an individual places on his or her career and work roles in comparison to other life roles, which may change over time for an individual (Rounds & Armstrong, 2005). Given the recent emphasis

on the psychology of working (Blustein, 2006), the notion of career salience may take on greater weight in examining of the meaning of work in individuals' lives.

Career beliefs and thoughts. Several researchers have focused on irrational beliefs or cognitions that interfere with career decision making or progress. For example, cognitive information-processing theory (Peterson, Sampson, Lenz, & Reardon, 2002) focuses on an individual's ability to process information effectively related to self-knowledge, occupational knowledge, decision-making skills, and executive processing (metacognitions about career decision cognitive strategies). Similarly, Krumboltz's (1996) social learning theory of career choice included cognitive interventions such as restructuring to address maladaptive career beliefs.

MEASURES OF CAREER DEVELOPMENT CONSTRUCTS

Psychometric/Theoretical Issues

Development of a psychological test, when done well and thoroughly, is a complex process with a number of iterative steps (Dawis, 1987; Walsh & Betz, 2001). It is beyond the scope of this chapter to fully describe the process of developing a psychometrically sound instrument, and readers are referred to the chapters in Part I of this volume for more information.

Some of the most difficult psychometric problems, identified early on by Super and Crites, relate to the developmental nature of the constructs. A characteristic of stage models of development is that prototypical behaviors vary by stage and that progression through stages is evinced through qualitative rather than quantitative differences. In other words, vocational/career maturity "looks" different at various ages/stages, rather than merely increasing in a linear fashion with age. For example, expansion of the number and range of careers under consideration may be an indicator of career maturity for a young adolescent, whereas reduction or narrowing of careers under consideration may be an indicator of career maturity for an older adolescent. How, then, does one create a measure to assess level of career maturity? Should there be different items, different scoring, and different norms, by stage of career development or by age?

Several instruments have been criticized for inadequate reliability (or at least some subscales suffer from internal inconsistency); reviewers and test developers have suggested that scales with these problems might be used as "checklists" or to stimulate discussion within counseling sessions. In other words, if scales are not used in selection or decision making, then lower reliability might be tolerated. For example, Krumboltz (1999) reported internal consistency coefficients for CBI scales as low as .16; Walsh and Betz (2001) concluded that most of the 25 CBI scales "would not qualify as 'scales' in the psychometric sense" (p. 302), noting that Krumboltz recommended using the instrument as a stimulus for discussion. Similarly, the ACCI is "most readily justified psychometrically . . . as a concerns checklist that can stimulate discussion," versus as a "scale" per se (Walsh & Betz, 2001, p. 295).

Many of the early measures of career maturity showed strong correlations with intellective variables (Betz, 1988; Westbrook, 1983), raising questions about construct and discriminant validity. A measure of any given construct should correlate more strongly with other measures of the same construct than it does with measures of separate constructs. Moreover, a measure of a construct should be predictive of something relevant and important. In the case of career maturity, measures should demonstrate theoretically consistent relationships to variables such as career satisfaction or coping ability.

Measures of other career development constructs, such as decision-making or dysfunctional thoughts, may have fewer psychometric challenges. When a construct does not have an explicit developmental component, there are fewer expectations about change with age that must be built into the measure. However, it is incumbent upon developers of inventories to identify clearly any developmental expectations that may be associated with a construct and to take care when writing items to avoid unintended developmental expectations.

Assessment of Career Maturity and Adaptability

As noted earlier, the progenitors of theory and research regarding vocational/career maturity—Super and Crites—each developed instruments to measure the construct, the CDI and CMI, respectively. These two measures were revised and used extensively by practitioners and researchers. They are discussed first, followed by three additional measures, the ACCI, the Career Mastery Inventory (CMAS; Crites, 1990), and the Career Attitudes and Strategies Inventory (CASI; Holland & Gottfredson, 1994).

The CDI (Super, Thompson, Lindeman, Jordaan, & Myers, 1979, 1981). The CDI first appeared in 1971 with three dimensions, the product of Super's ongoing work through the Career Pattern Study (Super et al., 1957; Super & Overstreet, 1960). The CDI was commercially published in 1981 by Consulting Psychologists Press and yielded eight scores (five separate scales and three composites). The current version of the CDI (available at http://www.vocopher.com) still comprises these scales. Part I, Career Orientation, includes Career Planning, Career Exploration, Career Decision Making, and World-of-Work Information. Three composite scores also are available: Career Development Attitudes (the sum of Career Planning and Career Exploration), Career Development Knowledge and Skills (the sum of Career Decision Making and World-of-Work Information), and a total score that is the sum of all four scales. Part II is Knowledge of Preferred Occupational Group, consisting of 40 items tapping knowledge about the respondent's self-identified occupational choice. There are two forms of the CDI, one for junior and senior high school students, and one for college/university students.

The CMI (Crites, 1974; Crites & Savickas, 1996). Crites published the Vocational Maturity Inventory (later renamed the CMI) in 1973, with a revision in 1978; this original 175-item measure is composed of the Attitude Scale with five scores and the Competence Test with five scores. A further revision (Crites & Savickas, 1996; available at http://www.vocopher.com) resulted in a 50-item inventory with two scales (Attitude and Competence). The CMI was designed to measure the two process-related aspects of career maturity in Crites's hierarchical model. The Competence Test (25 items) represents five general competencies: self-appraisal, occupational information, goal selection, planning, and problem solving. The Attitude Scale (25 items) represents the five domains of decisiveness, active involvement, independence in decision making, acceptance of certain realities about work, and ability to compromise.

The ACCI (Super et al., 1985). The ACCI measures career concerns that are typical of Super's stages of Exploration, Establishment, Maintenance, and Disengagement as well as consideration of career change. The ACCI evolved from an attempt to develop an adult version of the CDI, based on Super's (1977) model of career maturity in midcareer. The original intent was to develop a measure with the same dimensions as the CDI (i.e., planfulness, exploration, information, and decision making), but it proved difficult to write items for adults that adequately addressed these dimensions (Betz, 1988). Thus, only the planfulness dimension remained in the CDI-Adult and in the subsequent ACCI. The current version (available at http://www.vocopher.com) is a 60-item instrument that results in 12 scales, corresponding to the three major substages within each of the four stages: the Exploration substages of Crystallization, Specification and Implementation; the Establishment substages of Stabilizing, Consolidating, and Advancing; the Maintenance substages of Holding, Updating, and Innovating; and, the Disengagement substages of Deceleration, Retirement Planning, and Retirement Living. Several researchers have concluded that the ACCI is best considered a measure of current career concerns (vs. a measure of career development stage per se) or as a checklist of concerns to stimulate discussion within counseling (Cairo, Kritis, & Myers, 1996; Savickas, Passen, & Jarjoura, 1988; Walsh & Betz, 2001).

The CMAS (Crites, 1990). The CMAS was designed to assess career adjustment in adulthood, focusing on six developmental tasks in Super's Establishment stage and following Campbell and Cellini's (1981) diagnostic taxonomy of adult

career problems: Organizational Adaptability, Position Performance, Work Habits and Attitudes, Co-Worker Relationships, Advancement, and Career Choice and Plans. The CMAS (available at http://www.vocopher.com) consists of 90 true–false items, and norms are based on over 5,000 employed adults in a variety of industries.

The CASI (Holland & Gottfredson, 1994). The CASI consists of 130 items scored on nine scales measuring "work adaptation": Job Satisfaction, Interpersonal Abuse, Work Involvement, Family Commitment, Skill Development, Risk-Taking Style, Dominant Style, Geographical Barriers, and Career Worries; in addition, there is a checklist of 21 career obstacles. The CASI was designed to be self-administered and self-scored and is available from Psychological Assessment Resources.

Measures of Additional Career Development Constructs

Career Beliefs Inventory (CBI; Krumboltz, 1999). The CBI is a 96-item measure designed to identify irrational and illogical beliefs that individuals use in the process of making career decisions. The CBI yields 25 scales grouped under five general headings: (a) My Current Career Situation, (b) What Seems Necessary for My Happiness, (c) Factors That Influence My Decisions, (d) Changes I Am Willing to Make, and (e) Effort That I Am Willing to Initiate. The CBI is self-scorable and is available from Consulting Psychologists Press.

Career Decision-Making Difficulties Questionnaire (CDDQ; Gati et al., 1996). The CDDQ is a 34-item measure of concerns affecting career decisions, based on Gati's hierarchical model described earlier, resulting in 10 scales: Lack of Motivation, General Indecisiveness, Dysfunctional Beliefs, Lack of Information About the Career Decision-Making Process, Lack of Information About the Self, Lack of Information About Occupations, Lack of Information About Obtaining Information, Unreliable Information, Internal Conflicts, and External Conflicts. The measure is available online at no cost along with psychometric information and related measures (http://kivunim.huji.ac.il/cddq).

Career Decision-Making Profiles (CDMP; Gati et al., 2010). The CDMP is a 36-item measure of dimensions related to decision-making behavior, based on the Gati et al. (2010) model described earlier. Scores are available on 11 scales: Information Gathering, Information Processing, Locus of Control, Effort Invested in the Process, Procrastination, Speed of Making the Final Decision, Consultation With Others, Dependence on Others, Desire to Please Others, Striving Towards an "Ideal Occupation," and Willingness to Compromise. As with the CDDQ, the CDMP and supporting materials are available online at no cost (http://kivunim.huji.ac.il/cddq/cdmpinfo.htm).

Career Decision Scale (CDS; Osipow, Carney, Winer, Yanico, & Koschier, 1976). The CDS is an 18-item measure of career choice status and career-decision making difficulties. The Certainty scale comprises two items, and the Indecision scale comprises 16 items. The original intent of the measure was to indicate types of indecision; however, accumulated research leads to interpretation as a general measure of indecision (Osipow & Winer, 1996; Savickas, 2000). The CDS is available from Psychological Assessment Resources.

Career Decision Self-Efficacy Scale (CDSE; Taylor & Betz, 1983). The CDSE is a 50-item scale designed to measure self-efficacy beliefs about career decision making; a 25-item short form is also available (Betz & Taylor, 1994). Both forms yield five subscales: Accurate Self-Appraisal, Gathering Occupational Information, Goal Selection, Making Plans for the Future, and Problem Solving. Widely used as a research instrument that may also serve as a screening device in counseling settings, the CDSE is available from its author (N. Betz at Ohio State University).

Career Factors Inventory (CFI; Chartrand, Robbins, Morrill, & Boggs, 1990). The CFI is a 21-item measure "designed to help people determine whether they are ready to engage in the career decision-making process" (Chartrand & Robbins, 1997, p. 1). Scores are reported on four scales: Need for Career Information, Need for Knowledge, Career Choice Anxiety, and Generalized Indecisiveness.

The CFI is available from Consulting Psychologists Press.

Career Thoughts Inventory (CTI; Sampson, Peterson, Lenz, Reardon, & Saunders, 1996). The CTI is a 48-item inventory developed to measure dysfunctional thinking regarding career decision making; scores are reported on the three scales of Decision-Making Confusion, Commitment Anxiety, and External Conflict. The CTI is available from Psychological Assessment Resources.

Salience Inventory (Super & Neville, 1985). The Salience Inventory (available at http://www.vocopher.com) was designed to assess the importance of the work role relative to other roles in an individual's life. It consists of 170 items divided into three scales (Commitment, Participation, and Value Expectation), within each of five major life roles: homemaker, worker, student, citizen, and leisurite.

FUTURE DIRECTIONS IN THE ASSESSMENT OF CAREER DEVELOPMENT CONSTRUCTS

Renewed Attention and Emerging Attempts to Measure

Several recent, productive lines of research have brought renewed attention to the construct of career maturity, recast as career adaptability. An interesting feature of this work is that, in most cases, researchers are returning to earlier scales to operationalize career adaptability. For example, Hirschi (2009) used scales from the CDI and CMI as indicators of career adaptability; other researchers have used the CDSE (Duffy & Blustein, 2005) or the CDS (Creed, Fallon, & Hood, 2009). Perusal of recent literature suggests that the constructs are still valued, but relatively little attention has been paid to developing new measures or even refining existing measures. Some measures have had relatively little recent attention, even from their developers (e.g., Career Factors Inventory), whereas others have flourished despite lack of commercial publication (e.g., CDSE).

Although these new research directions hold promise, the question arises as to whether the extant scales have sufficient reliability and validity to use; some of the problems noted in earlier reviews have not been satisfactorily resolved. Moreover, the construct of career adaptability was intended to replace career maturity, yet the scales used to operationalize adaptability are essentially the scales designed to measure career maturity. Additionally, much of this research has been conducted outside of the United States, such as with Swiss high school students or Australian college students. Given that career maturity/adaptability could be influenced by societal and economic contexts, it is yet to be seen if this research translates to other settings.

One possible outcome of this renewed attention to the construct of career adaptability is the development of new instruments. An example is the Career Futures Inventory (CFI; Rottinghaus, Day, & Borgen, 2005), a 66-item measure of "positive career planning attitudes." Three of the seven CFI scales measure components of career adaptability: Career Optimism, Career Transition Confidence, and Control.

New models and methods of test delivery. Methods of test delivery (administering, scoring, and reporting) have changed tremendously since the heyday of career maturity measures. For example, Gati's CDDQ and CDMP are both administered and scored on the Internet, at no cost to the test taker or administrator. Furthermore, the website for the CDDQ and CDMP includes easily accessible psychometric information and research support for the instruments, by means of summaries and links to published articles. Such an arrangement offers several advantages to counselor and client: The inventories are free, portable, and accessible, and scoring and feedback is immediate. In comparison, common practice in earlier days of frequently used inventories (e.g., the CDS, ACCI, CDI, and CMI) required counselors to have expensive testing and scoring materials on hand, required control over administration and scoring, and often involved substantial turnaround time for score reports. If a primary purpose of these measures is to be diagnostic screening devices, or to be used directly in counseling sessions, then easier, cheaper, and faster are better.

Moreover, this model of delivery is an innovative approach to the distribution of psychological

assessment inventories. It is reminiscent of John Holland's "counselor-free" approach to interest testing (via the Self-Directed Search [SDS]), in that individuals may take the inventories and receive detailed feedback directly, without any counselor intervention. However, it goes beyond the SDS model by making the inventories free and offers a considerable amount of free technical support to counselors and other career development professionals. Such a model differs substantially from the traditional marketplace of the commercial test publisher.

Another sign of this new delivery model is evidenced at the website http://www.vocopher.com, established and administered by vocational psychology researchers. The CMI, CDI, ACCI, CMAS, and Salience Inventory are all now available at the website, because of the initiative of a few individuals and, presumably, because the copyrights on these inventories have been relinquished. Vocopher is a "career collaboratory" whose purpose "is to provide researchers and counselors with resources with which to further their research and assist their clients." The website would be strengthened by the addition of technical information for counselors and researchers.

The near demise of earlier measures of career maturity. If not for the Vocopher website, the old stalwarts of career maturity assessment—the CDI and CMI—would be unavailable to researchers and practitioners. Other measures/manuals, such as *My Vocational Situation* (Holland, Daiger, & Power, 1980) or the Assessment of Career Decision Making (Buck & Daniels, 1985; Harren, 1979) have essentially disappeared, out of print and no longer carried by their commercial publishers. One conclusion that might be drawn from the disappearance of these measures is that they have lost their utility or cachet. However, there also seems to be a resurgence of interest, certainly in the career adaptability portion. Moreover, measures of career development constructs seem to be on the upswing, as evidenced by the availability of measures such as the CDDQ and CDMP. The current state of affairs raises questions about theories of career development, the construct of career maturity, and the ability to measure readiness for career decision making or career adaptability.

Future of the Constructs

What is the future of career adaptability and other career development constructs? As noted by Crites (1961), the basic assumption underlying the construct of career maturity is that "vocational behavior changes systematically in certain ways with increasing age" (Crites, 1961, p. 255)—in the direction of becoming more goal directed, more realistic, and more independent. However, the world—and the nature of work itself—has changed substantially (Blustein, 2006; DeBell, 2006). A new direction in life-span psychology describes an "emerging adulthood" stage that occurs between adolescence and young adulthood (Arnett, 2000), which may serve to reinvigorate the discussion of career adaptability and career development. How much does the construct and measurement of career maturity and career development still offer today?

Furthermore, more construct explication is necessary. Is "career adaptability" a trait? Related to other traits? Can it be learned or improved? How does career maturity/adaptability fit into the bigger nomological net? There is implied *stability* of content constructs—interests, values, personality all are likely to remain relatively the same over time. However, there is implied *change* in constructs related to process, suggesting that more complex theoretical and psychometric approaches might be necessary.

Relatedly, what needs to be done about measuring career maturity and/or career adaptability? If the construct still has meaning and utility, then more attention needs to be directed to measurement issues. Constructing a measure that is developmental in nature brings a set of challenges: It takes a considerable amount of time; cross-sectional data are useful but not sufficient, and longitudinal data are necessary but not sufficient; any hypothesized/observed longitudinal changes could be due to historical effects as well as "true" changes; and changes in structure of work/careers may not be observed in data from scales. It may help to get away, both theoretically and psychometrically, from the idea that the construct is tied to

chronological age, an advantage of the transition to adaptability or "readiness."

Implications for Assessing Career Development

It may be useful to return to the discussion of career choice *process* versus career choice *content*. Assessment focused on content (interests, values, abilities, personality) is well established and supported by commercial publishers, making the assessment of these constructs easily accessible to counselors and clients. In contrast, assessment focused on the process of career choice is not, and so practitioners are unlikely to use formal standardized measures to determine a client's level of readiness for decision making or career adaptability. The current state of the content versus process distinction leads to the following conundrum. As Savickas (2000) aptly noted, "all too often career counselors have worked with clients only to find that they still cannot make a career choice" (p. 429), yet typically the assessment conducted as part of that career counseling is focused solely on career choice content. Although many career counselors may use informal methods of assessing clients' level of readiness, a reliable and valid instrument to use in diagnostic screening and subsequent treatment planning would be very helpful.

Another piece of the conundrum related to career choice process is that assessment for screening and treatment planning changes the counselor-client assessment dynamic outlined by Duckworth (1990), with testing done for the benefit of client and counselor and to generate information for both the client and counselor. Regarding career choice content, virtually all of the information generated through assessment is shared directly with the client, whereas that is less likely to occur for career adaptability or readiness.

In a world of shrinking budgets, it may become more difficult to justify process measures—indeed, to justify assessment in general. Clients may be willing to pay for assessment related to content (interests, values) but not process (decision-making style, career adaptability). In high school and college/university settings, where the majority of career counseling occurs, budget concerns may be particularly salient. The innovative forms of test delivery, and renewed attention to psychometric concerns, are welcome directions in the assessment of career development.

References

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist, 55*, 469–480. doi:10.1037/0003-066X.55.5.469

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. doi:10.1037/0033-295X.84.2.191

Bartlett, W. E. (1971). Vocational maturity: Its past, present, and future development. *Journal of Vocational Behavior, 1*, 217–229. doi:10.1016/0001-8791(71)90023-6

Betz, N. E. (1988). The assessment of career development and maturity. In W. B. Walsh & S. H. Osipow (Eds.), *Career decision making* (pp. 77–136). Hillsdale, NJ: Erlbaum.

Betz, N. E., & Taylor, K. M. (1994). *Career Decision-Making Self-Efficacy Scale manual*. Columbus: Ohio State University, Department of Psychology.

Blustein, D. L. (2006). *The psychology of working: A new perspective for career development, counseling, and public policy*. Mahwah, NJ: Erlbaum.

Buck, J. N., & Daniels, M. G. (1985). *Assessment of Career Decision Making (ACDM) manual*. Los Angeles, CA: Western Psychological Services.

Cairo, P. C., Kritis, K. J., & Myers, R. M. (1996). Career assessment and the Adult Career Concerns Inventory. *Journal of Career Assessment, 4*, 189–204. doi:10.1177/106907279600400205

Campbell, R. E., & Cellini, J. V. (1981). A diagnostic taxonomy of adult career problems. *Journal of Vocational Behavior, 19*, 175–190. doi:10.1016/0001-8791(81)90057-9

Chartrand, J. M., & Camp, C. C. (1991). Advances in the measurement of career development constructs: A 20-year review. *Journal of Vocational Behavior, 39*, 1–39. doi:10.1016/0001-8791(91)90002-4

Chartrand, J. M., & Robbins, S. B. (1997). *Career factors inventory*. Palo Alto, CA: Consulting Psychologists Press.

Chartrand, J. M., Robbins, S. B., Morrill, W. H., & Boggs, K. (1990). Development and validation of the Career Factors Inventory. *Journal of Counseling Psychology, 37*, 491–501. doi:10.1037/0022-0167.37.4.491

Creed, P. A., Fallon, T., & Hood, M. (2009). The relationship between career adaptability, person and situation variables, and career concerns in young

adults. *Journal of Vocational Behavior, 74*, 219–229. doi:10.1016/j.jvb.2008.12.004

Crites, J. O. (1961). A model for the measurement of vocational maturity. *Journal of Counseling Psychology, 8*, 255–259. doi:10.1037/h0048519

Crites, J. O. (1965). Measurement of vocational maturity in adolescence: I. Attitude test of the vocational development inventory. *Psychological Monographs: General and Applied, 79*(2), 36.

Crites, J. O. (1972). Career maturity. *Measurement in Education, 4*, 1–8.

Crites, J. O. (1974). Problems in the measurement of vocational maturity. *Journal of Vocational Behavior, 4*, 25–31. doi:10.1016/0001-8791(74)90088-8

Crites, J. O. (1990). *Career Mastery Inventory*. Boulder, CO: Crites Career Consultants.

Crites, J. O., & Savickas, M. L. (1996). Revision of the Career Maturity Inventory. *Journal of Career Assessment, 4*, 131–138. doi:10.1177/106907279600400202

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481–489. doi:10.1037/0022-0167.34.4.481

Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of Counseling Psychology, 39*, 7–19. doi:10.1037/0022-0167.39.1.7

DeBell, C. (2006). What all applied psychologists should know about work. *Professional Psychology: Research and Practice, 37*, 325–333. doi:10.1037/0735-7028.37.4.325

Duckworth, J. (1990). The counseling approach to the use of testing. *The Counseling Psychologist, 18*, 198–204. doi:10.1177/0011000090182002

Duffy, R. D., & Blustein, D. L. (2005). The relationship between spirituality, religiousness, and career adaptability. *Journal of Vocational Behavior, 67*, 429–440. doi:10.1016/j.jvb.2004.09.003

Gati, I., Krausz, M., & Osipow, S. H. (1996). A taxonomy of difficulties in career decision-making. *Journal of Counseling Psychology, 43*, 510–526. doi:10.1037/0022-0167.43.4.510

Gati, I., Landman, S., Davidovitch, S., Asulin-Peretz, L., & Gadassi, R. (2010). From career decision-making styles to career decision-making profiles: A multidimensional approach. *Journal of Vocational Behavior, 76*, 277–291. doi:10.1016/j.jvb.2009.11.001

Ginzberg, E., Ginsburg, S. W., Axelrad, S., & Herman, J. L. (1951). *Occupational choice*. New York, NY: Columbia University Press.

Gribbons, W. D., & Lohnes, P. R. (1968). *Emerging careers*. New York, NY: Teachers College Bureau.

Hansen, J. C. (2005). Assessment of interests. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 281–304). New York, NY: Wiley.

Harren, V. A. (1979). A model of career decision making for college students. *Journal of Vocational Behavior, 14*, 119–133. doi:10.1016/0001-8791(79)90065-4

Hirschi, A. (2009). Career adaptability development in adolescence: Multiple predictors and effect on sense of power and life satisfaction. *Journal of Vocational Behavior, 74*, 145–155. doi:10.1016/j.jvb.2009.01.002

Holland, J. L., Daiger, D. C., & Power, P. G. (1980). *My vocational situation*. Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L., & Gottfredson, G. D. (1994). *Manual for the Career Attitudes and Strategies Inventory*. Odessa, FL: Psychological Assessment Resources.

Jepsen, D. A., & Dilley, J. S. (1974). Vocational decision-making models: A review and comparative analysis. *Review of Educational Research, 44*, 331–349.

Kelly, K. R., & Lee, W. C. (2002). Mapping the domain of career decision problems. *Journal of Vocational Behavior, 61*, 302–326. doi:10.1006/jvbe.2001.1858

Koen, J., Klehe, U. -C., Van Vianen, A. E. M., Zikic, J., & Nauta, A. (2010). Job-search strategies and reemployment quality The impact of career adaptability. *Journal of Vocational Behavior, 77*, 126–139. doi:10.1016/j.jvb.2010.02.004

Krumboltz, J. D. (1996). A learning theory of career counseling. In M. L. Savickas & W. B. Walsh (Eds.), *Handbook of career counseling theory and practice* (pp. 55–80). Palo Alto, CA: Davies-Black.

Krumboltz, J. D. (1999). *Career Beliefs Inventory: Applications and technical guide* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Lonborg, S. D., & Hackett, G. (2006). Career assessment and counseling for women. In W. B. Walsh & M. J. Heppner (Eds.), *Handbook of career counseling for women* (2nd ed., pp. 103–166). Mahwah, NJ: Erlbaum.

Osipow, S. H., Carney, C. G., Winer, J., Yanico, B., & Koshier, M. (1976). *The career decision scale*. Odessa, FL: Psychological Assessment Resources.

Osipow, S. H., & Winer, J. L. (1996). The use of the Career Decision Scale in career assessment. *Journal of Career Assessment, 4*, 117–130. doi:10.1177/106907279600400201

Parsons, F. (1909). *Choosing a vocation*. Boston, MA: Houghton Mifflin.

Peterson, G. W., Sampson, J. P., Jr., Lenz, J. G., & Reardon, R. C. (2002). A cognitive information processing approach to career problem solving and decision making. In N D. Brown (Eds.), *Career choice and*

development (4th ed., pp. 312–369). San Francisco, CA: Jossey-Bass.

Phillips, S. D., & Jome, L. M. (2005). Vocational choices: What do we know? What do we need to know? In W. B. Walsh & M. L. Savickas (Eds.), *Handbook of vocational psychology* (3rd ed., pp. 127–153). Mahwah, NJ: Erlbaum.

Phillips, S. D., & Pazienza, N. J. (1988). History and theory of the assessment of career development and decision making. In W. B. Walsh & S. H. Osipow (Eds.), *Career decision making* (pp. 1–31). Hillsdale, NJ: Erlbaum.

Rottinghaus, P. J., Day, S. X., & Borgen, F. H. (2005). The Career Futures Inventory: A measure of career-related adaptability and optimism. *Journal of Career Assessment, 13*, 3–24. doi:10.1177/1069072704270271

Rounds, J. B., & Armstrong, P. I. (2005). Assessment of needs and values. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 305–329). New York, NY: Wiley.

Sampson, J. P., Jr., Peterson, G. W., Reardon, R. C., & Lenz, J. G. (2000). Using readiness assessment to improve career services: A cognitive information-processing approach. *Career Development Quarterly, 49*, 146–174. doi:10.1002/j.2161-0045.2000.tb00556.x

Sampson, J. P., Peterson, G. W., Lenz, J. G., Reardon, R. C., & Saunders, D. E. (1996). *Career Thoughts Inventory*. Odessa, FL: Psychological Assessment Resources.

Savickas, M. L. (1997). Career adaptability: An integrative construct for life-span, life-space theory. *Career Development Quarterly, 45*, 247–259. doi:10.1002/j.2161-0045.1997.tb00469.x

Savickas, M. L. (2000). Assessing career decision making. In C. E. Watkins & V. Campbell (Eds.), *Testing and assessment in counseling practice* (pp. 429–477). Mahwah, NJ: Erlbaum.

Savickas, M. L. (2002). Career construction: A developmental theory of vocational behavior. In D. Brown & Associates, *Career choice and development* (4th ed., pp. 149–205). San Francisco, CA: Jossey-Bass.

Savickas, M. L. (2005). The theory and practice of career construction. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 42–70). New York, NY: Wiley.

Savickas, M. L., Passen, A. J., & Jarjoura, D. G. (1988). Career concern and coping as indicators of adult vocational development. *Journal of Vocational Behavior, 33*, 82–98. doi:10.1016/0001-8791(88)90035-8

Sheppard, D. I. (1971). The measurement of vocational maturity in adults. *Journal of Vocational Behavior*, *1*, 399–406. doi:10.1016/0001-8791(71)90040-6

Slaney, R. B. (1988). The assessment of career decision making. In W. B. Walsh & S. H. Osipow (Eds.), *Career decision making* (pp. 33–76). Hillsdale, NJ: Erlbaum.

Super, D. E. (1953). A theory of vocational development. *American Psychologist, 8*, 185–190. doi:10.1037/h0056046

Super, D. E. (1955). Dimensions and measurement of vocational maturity. *Teachers College Record, 57*, 151–163.

Super, D. E. (1977). Vocational maturity in midcareer. *Vocational Guidance Quarterly, 25*, 294–302. doi:10.1002/j.2164-585X.1977.tb01242.x

Super, D. E. (1980). A life-space, life-span approach to career development. *Journal of Vocational Behavior, 16*, 282–298. doi:10.1016/0001-8791(80)90056-1

Super, D. E. (1983). Assessment in career guidance: Toward truly developmental counseling. *Personnel and Guidance Journal, 61*, 555–562. doi:10.1111/j.2164-4918.1983.tb00099.x

Super, D. E., Crites, J. O., Hummel, R. C., Moser, H. P., Overstreet, P. L., & Warnath, C. F. (1957). *Vocational development: A framework for research*. New York, NY: Bureau of Publications, Teachers College, Columbia University.

Super, D. E., & Knasel, E. G. (1981). Career development in adulthood: Some theoretical problems and a possible solution. *British Journal of Guidance and Counselling, 9*, 194–201.

Super, D. E., & Neville, D. D. (1985). *The salience inventory*. Palo Alto, CA: Consulting Psychologists Press.

Super, D. E., & Overstreet, P. L. (1960). *The vocational maturity of ninth-grade boys*. New York, NY: Teachers College Bureau of Publications.

Super, D. E., Thompson, A. S., Lindeman, R. H., Jordaan, J. P., & Myers, R. A. (1979). *Career Development Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Super, D. E., Thompson, A. S., Lindeman, R. H., Jordaan, J. P., & Myers, R. A. (1981). *Career Development Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Super, D. E., Thompson, A. S., Lindeman, R. H., Myers, R. A., & Jordaan, J. P. (1985). *Adult Career Concerns Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Swanson, J. L. (2012). Measurement and assessment. In E. M. Altmaier & J. C. Hansen (Eds.), *The Oxford handbook of counseling psychology* (pp. 208–236). New York, NY: Oxford University Press.

Swanson, J. L., & D'Achiardi, C. (2005). Beyond interests, needs/values, and abilities: Assessing other important career constructs over the life span. In S. D. Brown &

R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 353–381). New York, NY: Wiley.

Swanson, J. L., & Fouad, N. A. (2010). *Career theory and practice: Learning through cases* (2nd ed.). Thousand Oaks, CA: Sage.

Taylor, K. M., & Betz, N. E. (1983). Application of self-efficacy theory to the understanding and treatment of career indecision. *Journal of Vocational Behavior, 22*, 63–81. doi:10.1016/0001-8791(83)90006-4

Walsh, W. B., & Betz, N. E. (2001). *Tests and assessment*. Upper Saddle River, NJ: Prentice Hall.

Westbrook, B. E. (1983). Career maturity: The concept, the instruments, and the research. In W. B. Walsh & S. H. Osipow (Eds.), *Handbook of vocational psychology* (pp. 261–303). Hillsdale, NJ: Erlbaum.

Westbrook, B. E., Cutts, C. C., Madison, S. S., & Arcia, M. (1980). The validity of the Crites model of career maturity. *Journal of Vocational Behavior, 16*, 249–281. doi:10.1016/0001-8791(80)90055-X

CHAPTER 21

ASSESSMENT OF NEEDS AND VALUES

Melanie E. Leuty

Discussion of needs and values has extended across disciplines including psychology, sociology, business, and anthropology. Within psychology, attention to needs and values within the areas of personality, social, vocational, and developmental specialties can be found. Despite this widespread interest in needs and values, assessment of needs and values has lost favor over the past few decades as more attention has been devoted to the study of attitudes that are generally much easier to assess (Rokeach, 1973). The lack of research on values and needs has led to great misunderstanding of the definition and importance of these constructs. One objective of this chapter is to provide some much needed clarity to this field by reviewing the terminology, theories, measurement, and assessment of needs and values. Some additional avenues for scholarship also are suggested.

TERMINOLOGY

A basic understanding of the terminology is required before delving into the history and assessment of needs and values, especially given frequent confusion over their description. Early definitions of values suggested that values are a form of preference (Allport, 1961) or an evaluation of something as being desirable (Kluckhorn, 1951). Later, Rokeach (1973) added that values incorporate cognitive, affective, and behavioral components, meaning that preference, or value, can be expressed in any or all three of these areas. Super (1980) posited that values serve as a motivator for one's behavior, as individuals are motivated to seek ways to satisfy or fulfill their values. More recent discussion of values by Schwartz (1994) highlights that values transcend specific situations and generally serve to assist one in judging and justifying one's actions. Thus, although definitions vary slightly, Schwartz (1992, 1994) noted that most definitions contend that values (a) are beliefs, (b) relate to desirable end states or behaviors, (c) are consistent across situations, (d) guide choice and evaluation of behavior and events, and (e) are ordered by relative importance.

Similarly, the meaning ascribed to the concept of needs also has been unclear. Like values, needs have been defined as preferences or desires. Whereas values are assumed to be stable across situations, needs tend to be more specific to a given situation (Brown, 1996), and some theorists assume needs are a subordinate form of values (Rokeach, 1973; Super, 1973). Moreover, Rokeach (1973) indicated that needs are biologically based (e.g., needing food, shelter, or money), whereas values are more likely cognitive representations of those needs (security, affiliation, etc.) differentiating between the two. Super (1973) also assumed that needs were subordinate to values, with the process of satisfying needs leading to the development of values. Consequently, the terms *needs* and *values* have been used interchangeably (Maslow, 1954). In the vocational psychology literature, needs generally have been treated as facets of higher order work values (see, e.g., Lofquist & Dawis, 1978).

DOI: 10.1037/14048-021
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

THEORIES OF NEEDS AND VALUES

Theory development related to values has a long history dating back to Spranger (1928). As one of the first to classify values, Spranger identified six different values—theoretical, economic, aesthetic, social, political, and religious. Rokeach (1982) has argued that Spranger's classification is too general and that individuals who rank these values similarly may be very different. For example, two individuals may rank economic values as most important, yet one is very conservative with money whereas the other values money to support a lavish lifestyle. Despite this criticism, Spranger's classification of values has been a major influence in numerous theories.

Rokeach developed one of the most widely accepted theories of values. Describing values as enduring beliefs, Rokeach (1973) identified two categories of values: terminal and instrumental. Instrumental values refer to desirable modes of conduct, such as being honest. Terminal values, however, describe desirable endstates such as social recognition. Rokeach (1973) also further organized instrumental values into those that refer to moral values and those that refer to competence values. Additionally, terminal values are subdivided into personal values, such as salvation, and social values, such as world peace. As Rokeach posited, few values are needed to capture all values, and he specifies 18 instrumental and 18 terminal values.

More recently, Schwartz (1992) postulated that three universal requirements for human existence—biological needs, social interaction needs, and survival and welfare needs of groups—serve as the basis for values. Schwartz (1994) delineated 10 value types that capture these three universals. Schwartz's *Motivational Types of Values* (see Table 21.1) derived from Rokeach's value system, are conceptualized as the universal motivators that represent continuous, rather than discrete motivators, for behavior. Within each of Schwartz's 10 motivational values, more specific values define each domain. For example, the domain of power can be described by values of success, wealth, authority, and social power. Schwartz's theory articulates that these motivational type values form a circumplex model with values forming the shape of a circle where values adjacent to each other are more similar in nature than values across from each other (see Schwartz, 1994). Moreover, the 10 values can be summarized by four motivational themes; openness to change, self-enhancement, conservatism, and self-transcendence. These themes are considered to form two bipolar themes, openness to change versus conservatism and self-enhancement versus self-transcendence.

Theoretical attention to needs has primarily focused on the areas of motivation and personality. Most theories posit that needs direct behavior so that one's needs are satisfied. The discussion of needs appears to have originated with Murray's (1938) need-press theory. Influenced by Jung and psychoanalytic theory, Murray explicated that needs were conscious and unconscious motivators and that two classes of needs exist—viscerogenic and psychogenic. *Viscerogenic needs* refers to basic needs

TABLE 21.1

Motivational Themes, Types, and Values of Schwartz's Theory of Universal Values

Motivational theme	Motivational types	Values
Openness to change	Self-direction	Independence, creativity, exploration
	Stimulation	Excitement, novelty, challenge
	Hedonism	Pleasure, gratification
Self-enhancement	Achievement	Success, competence
	Power	Social status, prestige, dominance
Conservatism	Security	Safety, harmony, stability
	Conformity	Restraint
	Tradition	Respect, commitment, acceptance
Self-transcendence	Benevolence	Preserving and enhancing the welfare of one's group
	Universalism	Understanding, appreciation, tolerance, protection of other's welfare

required for survival such as the needs for air, water, urination, and harm avoidance. In all, 12 viscerogenic needs are specified. *Psychogenic need*s are psychological in nature and pertain to needs related to inanimate objects, ambition, status, power, aggression, inhibition, affection, and to share information, covered by 29 separate needs (see Murray, 1938, pp. 77–83, for descriptions of all needs).

Murray (1938) speculated that needs influence behavior and feelings. The environment, on the one hand, provides stimuli that either hinders or facilitates satisfying individuals' needs. This environmental component is referred to as press. The interaction between an individual trying to satisfy needs and the environmental conditions is then indicative of one's behavior. In essence, Murray's theory is a person–environment theory that emphasizes the fit between an individual's needs and the satisfaction of those needs by the current environment.

Another early conceptualization of needs that has continued to be popular is Maslow's (1943) hierarchy of needs. Maslow contended that physiological, safety, love, esteem, and self-actualization needs are universal to all humans. Needs are prioritized such that physiological and safety needs are more important than succeeding needs. Furthermore, as one level of needs is satisfied, the next need becomes more salient and motivates one to seek ways to satisfy this need. Maslow's theory has remained popular despite limited evidence supporting his model (Wahba & Bridwell, 1976).

In the work motivation literature, Maslow's theory influenced the development of other needs theories—for example, hygiene-motivator theory (Herzberg, Mausner, & Snyderman, 1959), the job characteristics theory (Hackman & Lawler, 1971; Hackman & Oldham, 1976) and the existence-relatedness-growth theory (Alderfer, 1969). Of these, the hygiene-motivator theory was most popular. It proposed that two types of needs, motivator and hygiene needs, influenced job satisfaction and dissatisfaction. Needs that promote satisfaction (e.g., achievement, recognition, challenge, responsibility, and advancement) are referred to as motivators. Conversely, hygiene needs (e.g., company policies, supervision, interpersonal relationships, physical working conditions, job security, benefits, and salary) are related to dissatisfaction if absent. Although at one time it received much attention, Miner (2006) suggested that the mixed results of research on this model led to it eventually being abandoned for other models.

The vocational psychology literature has provided rich discussion of values and needs as they relate to career choice and adjustment. Super's life-span, life-space theory included early mention of work values as being an important aspect of individuals' vocational traits (Super, 1953; 1980). As a developmental theory, Super's life-span, life-space theory emphasizes a longitudinal view of career development (Super, Savickas, & Super, 1996), where *life-span* refers to the lifetime of an individual, which comprises five developmental stages—growth, exploration, establishment, maintenance, and disengagement—that the individuals progress through, and *life-space* refers to the roles that one occupies in life. He defines eight major life roles for individuals: child, student, leisurite, citizen, worker, homemaker, spouse, and parent (see Super, 1980, for further explanation). These roles are expected to interact and these interactions can be both positive and negative and may influence one's career development and choice.

A main proposition of Super's theory is that people are inherently different in their skills, abilities, interests, personality, self-concepts, and values and that careers also are unique in their requirements. Super asserts that people then make career choices in light of their understanding of their own abilities, interests, values, and choices. This understanding is labeled *self-concept* by Super and is assumed to develop as individuals progress through the five stages of the life span. Values, in Super's theory, are an element of self-concept and, thus, are influential to career choice. Although, overall, he provides little elaboration on the role of work values in his model, Super (1970) has asserted that values serve to motivate individuals to seek out work environments that can satisfy individuals' work values. Research generally supports Super's life-span, life-space theory (see Swanson, 1992, for a review). However, the life-span theory has been criticized for being segmented (Brown, 1990).

The most popular vocational theory incorporating work values is the Theory of Work Adjustment

(TWA; Dawis & Lofquist, 1984). As a person–environment fit theory, TWA emphasizes that people possess certain attributes that are matched or mismatched with the requirements of the work environment. TWA mentions needs and abilities as particularly important attributes of individuals to help in career choice and adjustment. Needs are described as specific requirements that one needs to survive in life and these needs are satisfied by the work environment (Dawis, 1996). Abilities assist individuals in satisfying their needs. For example, an individual may possess strong verbal abilities that aid in fulfilling one's need for authority over others. TWA also posits that values are higher order constructs comprising needs and specifies six values—achievement, comfort, status, altruism, safety, and autonomy—that were derived from factor analysis of 20 work needs (Lofquist & Dawis, 1978). In TWA, the match between one's needs and values and the needs reinforced in the working environment are central to predicting job satisfaction. The TWA was developed empirically and, as such, has much research to support its propositions (see Dawis, 1996; Dawis & Lofquist, 1984).

Although it has received little attention, Duane Brown's new theory of career choice and development incorporates values. In Brown's values-based, holistic theory (Brown 1996, 2002; Brown & Crace, 1996), values are conceptualized as having behavioral, affective, and cognitive components that guide behavior with consideration of both cultural values and work values. Cultural values are described as values derived from cultural groups, such as values pertaining to time orientation, activity, self-control, and social relationships (Brown, 2002). Work and cultural values are seen as essential to career choice, satisfaction, and career success. Furthermore, Brown indicated that contextual factors, such as socioeconomic status, gender, minority status, and abilities impact the extent to which an individual can choose and advance a career based on values. Much like TWA, Brown's theory assumes that individuals' job satisfaction and tenure are related to the match between a person's attributes and the environment's requirements. No direct research empirically supports Brown's theory; instead, it gains support from the broader research on work values (Brown, 2002).

ASSESSMENT METHODS

Methodological approaches to assessing values and needs have been of central concern in the literature. The main issue is the way in which data are quantified. Popular methods of measurement include ranking, rating, and paired comparisons, with card sorts being less popular. Additionally, a more recent method of measurement, best–worst scaling, has been introduced in values assessment. The ranking of values data was made popular with the Rokeach Value Survey and the assumption that values are by nature hierarchical. Ranking of values items requires individuals to order a list of values from most to least important, so that a list of values are ranked overall. Proponents of ranking methods note its advantages (Miethe, 1985; Rokeach, 1973). For example, ranked procedures yield ipsative data meaning that it gathers information on importance of a value relative to other values for an individual versus a normative population. Rokeach (1973) described this process as being more accurate, as individuals tend to prioritize their values when in situations that elicit multiple values.

However, ranking methods have been highly criticized. As Hicks (1970) noted, purely ipsative measures, such as those that use rankings, can only legitimately be used for intraindividual comparisons. Therefore, ranked data are not appropriate for comparison between individuals. Moreover, ranked data are generally limited to nonparametric statistical analyses which greatly limit the usefulness of the data. Others have noted that another drawback to ranking values is that ranking can be a difficult and time-consuming task for individuals (Alwin & Krosnick, 1985; Ovadia, 2004), which Rokeach (1973) admitted was an issue for the Value Survey, which contains two sets of 18 values ranked by test takers. Ranking also requires individuals to place one value as more important to another when no difference in importance may exist (McCarty & Shrum, 2000).

An alternative method of assessing values is rating, which has been posited as the most preferred method for assessing values (Krosnick & Alwin,

1989). Rating of values tends to include using a Likert-type scale to measure the amount of preference for a given value item, which allows for multiple values to be rated as equally important. This method produces normative data that can be used to make comparisons between individuals. Furthermore, many statistical tests can be used to analyze the data. Rating methods also may reduce participant burden and be faster to complete than other methods (McIntyre & Ryans, 1977).

Rating methods are not without their own problems. The most common issues with rating scales are their vulnerability to response bias. For instance, Schwartz, Verkasalo, Antonovsky, and Sagiv (1997) found a significant relationship between values and social desirability, suggestive of some values being overly endorsed due to perceptions of what society values versus what an individual values. Furthermore, given that values are generally represented as positive ideals, a major issue in using a rating method to assess values and needs is the tendency for individuals to endorse all items as being important, creating acquiescence bias (Schwartz & Bardi, 2001). This results in response bias when individuals fail to differentiate between items and respond to items in a generally favorable manner (McCarty & Shrum, 2000; see also Chapter 11, this volume). Some studies have found average of 30% of values rated equally (Maio, Roese, Seligman, & Katz, 1996) to upward of 60% of values rated equally in nearly half of a sample (Krosnick & Alwin, 1988), suggesting that acquiescence is a widespread problem with rated value scales.

Lee and colleagues (Lee, Soutar, & Louviere, 2007) noted that the effect of acquiescence may be stronger in different cultures which leads to some cultures scoring higher on values scales than do others. Differences in the selection of the midpoint or extremes on a scale also have shown to vary across cultures which can create further response bias (see Lee et al., 2007, for a brief review). This issue makes it difficult to ascertain what values are truly important. As well, this factor can falsely inflate correlations between values, making accurate conclusions from the data difficult.

Research comparing ranking and rating of values suggests that ranking may be superior (Krosnick & Alwin, 1988; Meithe, 1985; Rankin & Grube, 1980). However, others have commented that the research questions should guide the type of measurement because differences in reliability and validity are slight (Munson & McIntyre, 1979; Rankin & Grube, 1980; Thompson, Levitov, & Miederhoff, 1982) and rating versions actually may have slightly better evidence of predictive validity (Maio, Roese, Seligman, & Katz, 1996; Rankin & Grube, 1980). Because of the ease of use and production of data amenable to more statistical procedures, rating versions of values assessments have tended to be more popular (McCarty & Shrum, 2000).

Ovadia (2004) has suggested that ranking and rating systems do not have to be mutually exclusive. Combining both a ranked and rated approach to values measurement, according to Ovadia, can lead to deeper understanding of individuals' value systems and address limitations of both ranked and rated data. This type of measurement approach also may solve issues with poor differentiation of items when using a rated procedure. As McCarty and Shrum (2000) noticed in their work, a rank then rate procedure caused individuals to begin to anchor their responses and led to fewer issues with response bias. Additionally, as data are rated, a variety of statistical procedures can still be used with the data. Despite the advantages of this combined system of measurement of values, this method can be overly time consuming (McCarty & Shrum, 2000).

Although it may provide a solution to the issue of ranking versus rating values data, the paired-comparison method for values measurement has not been widely popular. Introduced by Thurstone (1927, 1954), the paired-comparison method, an alternative ranking procedure, involves choosing a preferred value in pairs of value statements. Each item of a measure is presented with every other item. The advantage of this procedure is that it allows differentiation of the relative importance of a value statement without producing data with limited statistical uses like that obtained from a ranking procedure. It also avoids the vulnerability of a rating method that may yield a fixed response style. Another advantage of a paired-comparison method is that the circularity of responses can be determined (Kendall & Babington Smith, 1940; Thurstone,

1927) to determine consistency in responding. For example, if Value A is preferred over Value B, and Value B is preferred over Value C, it is expected that Value A would be rated as preferred over Value C when compared. When this assumption is violated, it suggests that individuals are not either attending to items or are not reliable in their judgments (Gay, Weiss, Hendel, Dawis, & Lofquist, 1971; Kendall & Babington Smith, 1940). Thus, this method provides a way to examine the reliability of individuals responding by allowing for an index of consistency to be calculated. The largest drawback to this method, however, is that it requires that all item combinations be presented to the individual. Although this may not be an issue with smaller measures, larger lists (e.g., more than 20 items) become impractical as time and participant fatigue can interfere with data collection (Dawis, 1987).

The newest evolution in values measurement, best–worst scaling (also known as *maximum difference scaling*), has been promoted as overcoming the drawbacks of ranking, rating, and paired methods (Lee et al., 2007). First described by Louviere and Woodworth (1990) and Finn and Louviere (1992), best–worst scaling is an extension of Thurstone's (1927) concept of paired comparisons. Best–worst scaling has individuals choose the best and worst (i.e., most and least) options in a block of items. Advantages to this approach, similar to paired comparison methods, are that best–worst scaling can increase differentiation between items versus a rated procedure as well as provide a relatively easier cognitive task than overall ranking procedures because individuals are only asked to determine the least and most preferable options in a small set of items. Statistical packages, such as the MaxDiff program (http://www.sawtoothsoftware.com), are becoming available to make administration and analyses more feasible for professionals. Additionally, Marley and Louviere (2005) discussed the methodology behind using a best–worst scaling procedure to aid in data analysis.

In a comparison of best–worst scaling, rating, and ranking procedures with the List of Values (Kahle, 1983), Lee et al. (2007) found the best–worst scaling method outperformed other methods. Results suggested that best–worst scaling produced less skewed data, more meaningful intercorrelations between values, and more discriminating results correlated to behavioral items related to the values assessed. Similar results were found comparing a rating and best–worst scaled version of the Schwartz Value Survey (Lee, Soutar, & Louviere, 2008). Overall, they assert that in the measurement of values, best–worst scaling can produce more accurate information. Furthermore, as participants are not asked to assign numerical values to items, this method may be less susceptible to responses bias as a result of cultural differences.

Finally, a card sort procedure for assessing values also has been used. In its simplest form, card sorts have individuals distribute a list of values, each written on a separate card, in to piles ranging from most to least important. Slightly more sophisticated card sorts require individuals to sort cards into a specific number of categories, and may include guidelines of how many cards to sort into each group, to force the respondent into creating a normal distribution of responses. The Work Importance Locator (WIL; U.S. Department of Labor, 2000a), for instance, has individuals sort 20 cards, each with one value, into five categories ranging from least to most importance with four cards per category. Forcing responses into a set number per category aids in increasing variability and helps prevent response biases such as yea-saying or clustering responses around the mean.

When the procedures for a card sort are standardized, as described, the method is known as a Q-sort technique (Stephenson, 1953). This method is popular in personality assessment but has extended to use with values and needs assessment. As Dawis (1987) noted, Q-sort data can be used for different analyses like analysis of variance, correlation, and factor analysis when the data are not forced into a distribution. If data are forced into a distribution (e.g., a specific number of responses needed per category) the resulting data are ipsative and distributed around an individual's mean, making fewer analyses appropriate for use.

Suggestions for using values card sorts and other measures are widely available on the Internet, making this method very accessible (e.g., Knowdell, 1998; Miller, C'de Baca, Matthews, & Wilbourne,

2001; U.S. Department of Labor, 2000a). Although card sorts are very amenable to use in counseling situations, for interindividual comparisons, obtaining measurable data using card sorts can be more cumbersome. However, some statistical packages are available to analyze Q-sort data.

In summary, numerous methods for assessing needs and values have been used. The oldest approach is ranking, whereas newer methods, such as best–worst scaling, are still evolving to address the shortcomings of past strategies. Ranking and Q-sorts are amenable to use with individuals in a counseling setting, as they may assist individuals in determining what needs or values are most important relative to other values. These methods can be used for research, but their limits on the variety of statistical procedures make them less desirable for research in general. Rating methods tend to be a more popular method of measurement for research settings because they allow easy comparisons between individuals. However, they can be vulnerable to response bias when assessing needs and values. Paired-comparison methodology can overcome issues with response bias and may be easier for individuals to complete but also can be unfeasible when there are a large number of items. Best–worst scaling appears to provide many advantages, but it is a relatively new methodology for value and need assessment. More research on this approach may help determine its overall usefulness for assessing needs and values.

ESTABLISHED MEASURES OF VALUES AND NEEDS

Study of Values (SOV)

The SOV (Vernon & Allport, 1931) is one of the oldest measures of values and was the third most popular nonprojective assessment by 1970 until it eventually became obsolete by the 1990s, likely because of few modifications to keep item wording from becoming archaic (Kopelman, Rovenpor, & Guan, 2003). Based on Springer's six values (theoretical, economic, aesthetic, social, political, and religious), the SOV contains 45 items. The first 30 items are yes/no questions about one's preferences. The last 15 questions present scenarios where the individual can choose among four different responses based on his or her importance of different values. Kopelman and colleagues (2003) suggested that the SOV is unique and more advantageous from other values measures because its items are behaviorally focused (e.g., asking what action one would take in a given situation). The measure was updated in 1951 (Allport, Vernon, & Lindzey, 1951) and again in 1995 (found in Kopelman et al., 2003) to make the wording of the instrument more appropriate (e.g., updating *men* to *people*, including more recent cultural references).

Research on the updated version of the SOV suggests adequate evidence of reliability with internal consistency estimates (coefficient alpha) ranging from .55 (political) to .80 (religious) for the values scales (Kopelman et al., 2003). Furthermore, although evidence of validity has not been published on the updated version of the SOV, earlier versions were shown to relate to occupational choice, value changes, and interest measures (Allport, Vernon, & Lindzey, 1970). Given evidence of psychometric adequacy and amenable use in classroom and counseling settings, Kopelman et al. (2003) suggested that the SOV should be reconsidered as a viable values instrument.

Rokeach Values Survey (RVS)

Designed to capture universal values, the RVS contains 36 values that are ranked in importance by the test taker. The assessment is divided into two parts, the first part containing 18 terminal values and the second part containing 18 instrumental values. Earlier versions of the assessment included 12 items for each part but were later expanded to be more comprehensive (Rokeach, 1973).

Terminal values were generated from a review of the literature, and values obtained from a small sample of graduate students in psychology and adults from a Midwestern community (Rokeach, 1973). The process of developing instrumental values items differed. From a pool of 555 personality–trait words generated by Anderson (1968), items were selected for inclusion in the measure by retaining those perceived as widely applicable, discriminating across demographic groups, and not vulnerable to social desirability response bias.

The RVS has gone through several revisions. Most changes involved the presentation of the items. Initial versions (Forms A and B) included ranking of value items. Form C had participants sort values into three categories of importance (high, middle, and low) and then rank items within each category. Later versions (Forms D and E) returned to ranking items.

Currently, the RVS is available in two response formats, a ranked version and a rated version. Evidence collected by Thompson, Levitov, and Miederhoff (1982) has suggested that compelling evidence is available for the construct validity of the rated version of the RVS as well as the advantage of allowing comparisons across individuals for use in research. Moreover, Maio, Roese, Seligman, and Katz (1996) found that more evidence of predictive validity exists for the rated version, as did Rankin and Grube (1980). On the other hand, responses obtained from the ranked data are preferable when the purpose is to understand one individual's preferences, as the data is ipsative. Rankin and Grube (1980) contend that empirical evidence suggests that the test–retest reliability and convergent and discriminant validity for the ranked and rated version are equivalent.

Studies on the reliability and validity of the ranked RVS have been summarized by Rokeach (1973). Three week test–retest reliabilities of the RVS using samples of middle school, high school, and college students range from .62 to .74 for terminal values and .53 to .71 for instrumental values. Test–retest reliabilities over a 14 to 16 month period for a sample of college students were .69 for terminal values and .61 for instrumental values. Results suggest terminal values have slightly better retest reliability.

Although the RVS has been widely used, it has been criticized for little evidence of content and construct validity. Numerous authors have concluded that the RVS does not likely cover the breadth of all human values (Braithwaite & Law, 1985; Jones, Sensenig, & Ashmore, 1978; Kitwood & Smithers, 1975). Heath and Fogel (1978) examined the organization of values into terminal and instrumental categories and found that this organization is arbitrary given evidence that eight factors accounted for the data rather than two.

Schwartz Value Survey (SVS)

Developed from his theory of values, the SVS contains 57 value items (Schwartz, 1992, 1994), which assess 10 value domains (achievement, benevolence, conformity, hedonism, power, security, self-direction, stimulation, tradition, and universalism). The SVS includes items from Rokeach's measure and some additional items developed by Schwartz. Forms of the SVS have used both a ranked method and a rated method to assess values. On the rated version, values are measured using a rating scale ranging from −1 (*opposed to my principles*) to 0 (*not important*) to 7 (*of supreme importance*; Schwartz & Boehnke, 2004). A version using best–worst scaling has recently been introduced, and preliminary evidence suggests it approximates Schwartz's circumplex model better than a ranked version and is quicker to complete (Lee et al., 2008). A short version with just the 10 value domains, described by the corresponding 57 values, is also available. Preliminary research has found that the shortened version approximates the circumplex structure, similar to the original version, with the advantage of being much briefer (Lindeman & Verkasalo, 2005).

Research on Schwartz's theory has reinvigorated research on values with research conducted with samples worldwide. Two-dimensional methods (e.g., multidimensional scaling and similarly structured analysis) have supported the separation of values into the 10 types and the ordering of the data into a circumplex model (Schwartz, 1992, 1994). Analysis of the structure of the instrument using 21 samples from different countries found similar factor structure and scale equivalence (i.e., scores have similar meaning between groups) across cultures (Spini, 2003).

Minnesota Importance Questionnaire (MIQ)

As mentioned earlier, much of the research on values and needs has been completed as it pertains to a work. This focus being the case, some of the more popular measures of values are specific to work values and needs. The MIQ (Rounds, Henley, Dawis, Lofquist, & Weiss, 1981) is likely one of the most popular measures of work values and needs because the MIQ is more comprehensive than other work

values measures (Rounds, 1990). The MIQ measures 20 work needs, drawn largely from Schaffer's (1953) consolidation of 12 needs drawn from Murray's larger list of needs. As seen in Table 21.2, needs listed on the MIQ are organized into six separate values—achievement, comfort, status, altruism, safety, and autonomy.

The MIQ was first developed in a Likert-type format with 100 items that, despite demonstrating acceptable evidence of reliability and validity, elicited response bias with most items being positively endorsed (Gay, Weiss, Hendel, Dawis, & Lofquist, 1971). Because of this limitation, a paired-comparison version was created using the 20 items from the Likert-type version that had the highest correlation with their respective scale. This resulted in 190 pairs from the possible combinations of the 20 need items. In addition to reducing response bias by switching to a paired-comparison version, an index of circularity of responses was able to be included to provide data on random responding. An equivalent version that has respondents rank sets of five needs also is available (Rounds et al., 1981). This version does include one more need item (autonomy) that was included to balance the multiple rank-order format.

Ample psychometric evidence of reliability has been found for the paired-comparison version of the MIQ. Hendel and Weiss (1970) found that median scores for internal consistencies of scales (Hoyt coefficients) ranged from .77 to .81 and test–retest reliabilities for MIQ values scales have been found at .89 for immediate retesting and median scores for scales ranging from .46 to .79 over a period of 10 months. Hendel and Weiss (1970) examined the stability of individuals' profiles over time and found that the median stability of scores was .95 for immediate retesting, .75 over a 4-month interval, and .53 over a 10-month interval suggesting that the overall profile is fairly stable for up to a year.

Evidence of validity supports psychometric soundness for MIQ scores. Lofquist and Dawis (1978) suggested that the needs on the MIQ could be summarized by six overarching values based on factor analyses conducted using MIQ scores from a sample of 5,358 individuals, including college students, vocational rehabilitation clients, and employed workers, (data from Gay et al., 1971) and replicated with a sample of 3,283 vocational rehabilitation clients (data from Seaburg, Rounds, Dawis, & Lofquist, 1976). Other evidence suggests that MIQ scores demonstrate evidence of discriminate validity given low correlations (less than .30) with an ability measure (Weiss, Dawis, Lofquist, & England, 1966). Additionally, MIQ scores have been found to differentiate between different occupational groups (Weiss et al., 1966).

TABLE 21.2

Minnesota Importance Questionnaire Values and Needs

Work value	Needs
Achievement	Ability utilization, achievement
Comfort	Activity, independence, variety, compensation, security, work conditions
Status	Advancement, recognition, authority, social status
Altruism	Coworkers, social service, moral values
Safety	Company policies, supervision-human, supervision-technical
Autonomy	Creativity, responsibility

Ronen's Taxonomy of Needs

Although Ronen's taxonomy of needs (Ronen, Kraut, Lingoes, Aranya, 1979) has not been widely used, it provides an alternative assessment of vocational needs. Ronen's taxonomy was created in an effort to incorporate the need classifications of Herzberg (Herzberg et al., 1959), Alderfer (1969), and Maslow (1954), using the 14 need items from earlier work of Hofstede (Hofstede, Kraut, & Simonetti, 1977). The 14 needs included (advancement, area, autonomy, benefits, challenge, coworkers, earnings, manager, physical, recognition, security, skills, time, and training) are rated on their importance in an ideal job on a 5-point scale ranging from 1 (*utmost importance*) to 5 (*very little or no importance*) using one item to assess each need.

Research by Kraut and Ronen (1975) examined evidence of validity for the needs. Using samples of 2,376 individuals employed in sales and 6,331 repairpersons from an international organization located in five countries (Canada, France, Germany,

Japan, and the United Kingdom), needs were found to be ordered similarly across countries and predictive of job satisfaction for both salespersons ($r = .61$) and repairpersons ($r = .58$). Multidimensional scaling of the 14 needs with a sample of 2,600 sales and repairmen from a German company found that the spatial arrangement of the needs were arranged consistent with Maslow's hierarchy (Ronen et al., 1979). For instance, needs were ordered such that physiological needs (e.g., time, physical) were closely located to safety needs (e.g., security, earnings), which were located next to love (e.g., coworkers, mangers) and esteem needs (e.g., recognition). Self-actualization needs (e.g., advancement, training, skills, autonomy, and challenge) were farthest from physiological needs, as anticipated.

Super's Work Values Inventory

Super's measure of work values was created as part of the Career Pattern Study (Super, 1985) which followed the career development of a sample of ninth-grade boys for over 20 years and led to the development of Super's life-span, life-space theory. According to Super (1970), items for the Work Values Scale were selected from Spranger's theory (1928), the Allport–Vernon–Lindzey Study of Values (Allport, Vernon, & Lindzey, 1960), and research on job satisfaction and morale by Hoppock (1935) and Centers (1948). Some additional items were included based on the work of Darley and Hagenah (1955), Fryer (1931), Ginzberg and colleagues (Ginzberg, Ginsburg, Axelrad, & Herma, 1951), and Super (1957).

The most recent version is Super's Work Values Inventory—Revised (SWVI–R; Zytowski, 2006) which is based on a revision of the 1970 version (Super, 1970). Revisions included the deletion of three scales (Altruism, Esthetics, and Management) that were highly correlated with the content of other career measures, demonstrative of unsatisfactory evidence of discriminant validity. The revised version now has 12 work values scales—Achievement, Co-Workers, Creativity, Income, Independence, Lifestyle, Challenge, Prestige, Security, Supervision, Variety, and Workplace—assessed with six items each, resulting in a total of 72 items. Responses on the SWVI–R are rated on a 5-point scale ranging from 1 (*not important at all/not a factor in my job selection*) to 5 (*crucial/I would not consider a job without it*).

Because of the newness of the revised version, little research has been conducted to examine the psychometric properties of the SWVI–R. Robinson and Betz (2008), however, offer some preliminary evidence of reliability and validity beyond what is reported in the manual that mostly pertains to an earlier version of the SWV (Zytowski, 2006). Internal consistency estimates (coefficient alpha) for the 12 values scales ranged from .72 (Independence) to .88 (Income) in a sample of 426 university students. Robinson and Betz also examined intercorrelations between values scales which offered some evidence of discriminate and convergent validity, in other words scales with similar content shared higher correlations than did dissimilar scales. For instance Prestige had a large correlation with Achievement ($r = .67$), and Variety with Mental Challenge ($r = .70$). Finally, a factor analysis performed on the data found four factors (Environment, Esteem, Excitement, and Safety) fit the data, explaining 76% of the total variance in SWVI–R scores.

The Work Importance Locator and Work Importance Profiler

The U.S. government has more recently developed two work need and value assessments that are based on the multiple rank version of the MIQ. One is a multiple-rank version that is computerized (the Work Importance Profiler [WIP]), and the other is a shorter card sort with paper-and-pencil scoring (the WIL). The WIL (U.S. Department of Labor, 2000a) and the WIP (U.S. Department of Labor, 2000b) include assessment of work needs that are mostly identical to those on the MIQ, although some items have been slightly reworded to improve clarity and update language. The WIP contains all 21 needs from the MIQ multiple-rank version. However, the WIL was slightly modified for use. The social status need, originally part of the Recognition value scale, was deleted from the WIL for two reasons: The social status need did not provide added benefit to the recognition value, which includes four other needs; and eliminating the social status need simplified the self-scoring process (McCloy, Waugh,

Medsker, Wall, Rivkin, & Lewis, 1999a). Similar to the MIQ, the WIL and WIP organize work needs into six overarching values—achievement, independence, recognition, relationships, support, and working conditions.

The WIL is presented as a Q-sort task requiring individuals sort cards with each of the 20 values the cards into five categories from most to least important and assign exactly four cards to each category. After completing this, individuals are instructed on how to score their responses. This version is very amenable to use in an individual or group counseling situation. The WIP is a computerized assessment that is presented in a multiple-rank format like the MIQ. Although both are available on the Internet (http://www.onetcenter.org), the WIP is likely easier for individuals to administer and interpret on their own. Examination of the reliability and validity of the WIL and the WIP suggest that the WIP has higher test–retest reliability, higher estimates of internal consistency, and larger correlations with MIQ scales (McCloy et al., 1999a, 1999b). These results may be due to the ipsative nature of a card sorting task with the WIL (McCloy et al., 1999a). Thus, psychometric evidence supports the use of the WIP over the WIL.

FUTURE DIRECTIONS

Attention to understanding and assessing needs and values gained popularity during the middle of the century, but interest began to decline in the late 1970s. In the past few years, discussions of values and needs have increased, especially in business publications, suggesting that the study of values and needs will again be in style. As more individuals become interested in this topic, there are some areas in particular that provide opportunities to expand knowledge on needs and values.

First, a main issue in the science on needs and values is the diversity of theories and assessments in this area. Although only the most often used assessment instruments were highlighted here, the plethora of need and value instruments has made it difficult for the literature to progress (Roe & Ester, 1999). Additionally, some have suggested that the construct of work values has not been fully explored and additional values relevant to work situations may be missing from existing assessments (Berings, de Fruyt, & Bouwen, 2004; Rounds & Armstrong, 2005). Initial efforts have tried to compare and further our knowledge of the domains represented in the construct of work values (see Leuty & Hansen, 2011; Macnab & Fitzsimmons, 1987), but efforts focused on values in general have not been identified. Comparisons of different value and need instruments can examine evidence of construct validity, provide insight into the domains relevant to these constructs, and provide structure to theory development in this area.

Developing the literature on the use of best–worst scaling (Finn & Louviere, 1992; Louviere & Woodworth, 1990), a newer methodology in values research, could be used to explore further whether best–worst scaling is advantageous over other methods and invigorate the use of value and need assessments. Moreover, research that can provide more evidence for use of particular assessments in different settings (hiring decisions, therapy, career counseling, etc.) would be beneficial.

Finally, consideration of the influence of culture on need and value theory and assessment is an avenue for future study. Some theories, such as Super's and Schwartz's, have demonstrated their cross-cultural applicability, and, as such, expanded their evidence of validity and visibility. Examining other assessments and theories with differing cultures can provide important information for organizations and practitioners as the diversity of the U.S. population and business increases and demand for culturally applicable assessment and theory increases.

References

Alderfer, C. P. (1969). An empirical test of a new theory of human needs. *Organizational Behavior and Human Performance, 4*, 142–175. doi:10.1016/0030-5073(69)90004-X

Allport, G. W. (1961). *Pattern and growth in personality*. New York, NY: Holt, Rinehart & Winston.

Allport, G. W., Vernon, P. E., & Lindzey, G. (1951). *Manual to the study of values*. New York, NY: Houghton Mifflin.

Allport, G. W., Vernon, P. E., & Lindzey, G. A. (1960). *A study of values* (rev. ed.). Boston, MA: Houghton Mifflin.

Allport, G. W., Vernon, P. E., & Lindzey, G. A. (1970). *Study of values* (3rd ed.). Chicago, IL: Riverside.

Alwin, D. R., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49*, 535–552. doi:10.1086/268949

Anderson, N. H. (1968). Likableness ratings of 55 personality–trait words. *Journal of Personality and Social Psychology, 9*, 272–279. doi:10.1037/h0025907

Berings, D., de Fruyt, F., & Bouwen, R. (2004). Work values and personality traits as predictors of enterprising and social vocational interests. *Personality and Individual Differences, 36*, 349–364. doi:10.1016/S0191-8869(03)00101-6

Braithwaite, V. A., & Law, H. G. (1985). Structure of human values: Testing the adequacy of the Rokeach Values Survey. *Journal of Personality and Social Psychology Journal, 49*, 250–263. doi:10.1037/0022-3514.49.1.250

Brown, D. (1990). Summary, comparison, and critique of the major theories. In D. Brown, L. Brooks, & Associates. (Eds.), *Career choice and development* (2nd ed., pp. 338–363). San Francisco, CA: Jossey-Bass.

Brown, D. (1996). Brown's value-based holistic model of career and life-role choices and satisfaction. In D. Brown, L. Brooks, & Associates. (Eds.), *Career choice and development* (3rd ed., pp. 337–372). San Francisco, CA: Jossey-Bass.

Brown, D., & Crace, R. K. (1996). Values in life roles and outcomes: A conceptual model. *Career Development Quarterly, 44*, 211–223. doi:10.1002/j.2161-0045.1996.tb00252.x

Brown, D. E. (2002). The role of work and cultural values in occupational choice, satisfaction, and success: A theoretical statement. *Journal of Counseling and Development, 80*, 48–56. doi:10.1002/j.1556-6678.2002.tb00165.x

Centers, R. (1948). Motivational aspects of occupational stratification. *Journal of Social Psychology, 28*, 187–217. doi:10.1080/00224545.1948.9921768

Darley, J. G., & Hagenah, T. (1955). *Vocational interest measurement: Theory and practice*. Minneapolis: University of Minnesota Press.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481–489. doi:10.1037/0022-0167.34.4.481

Dawis, R. V. (1996). The theory of work adjustment and person-environment correspondence counseling. In D. Brown, L. Brooks, & Associates. (Eds.), *Career choice and development* (3rd ed., pp. 75–120). San Francisco, CA: Jossey-Bass.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment*. Minneapolis: University of Minnesota Press.

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing, 11*, 12–25.

Fryer, D. (1931). *The measurement of interests in relation to human adjustment*. New York, NY: Holt.

Gay, E. G., Weiss, D. J., Hendel, D. D., Dawis, R. V., & Lofquist, L. H. (1971). Manual for the Minnesota Importance Questionnaire. *Minnesota Studies in Vocational Rehabilitation* (No. 54).

Ginzberg, E., Ginsburg, S., Axelrad, S., & Herma, J. (1951). *Occupational choice*. New York, NY: Columbia University Press.

Hackman, J. R., & Lawler, E. E. (1971). Employee reactions to job characteristics. *Journal of Applied Psychology, 55*, 259–286. doi:10.1037/h0031152

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: A test of a theory. *Organizational Behavior and Human Performance, 16*, 250–279. doi:10.1016/0030-5073(76)90016-7

Heath, R. L., & Fogel, D. S. (1978). Terminal or instrumental? An inquiry into Rokeach's value survey. *Psychological Reports, 42*, 1147–1154. doi:10.2466/pr0.1978.42.3c.1147

Hendel, D. D., & Weiss, D. J. (1970). Individual inconsistency and reliability of measurement. *Educational and Psychological Measurement, 30*, 579–593.

Herzberg, F., Mausner, B., & Snyderman, B. (1959). *The motivation to work*. New York, NY: Wiley.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167–184. doi:10.1037/h0029780

Hofstede, G. H., Kraut, A. I., & Simonetti, S. H. (1977). Development of a core attitude survey questionnaire for international use. *JSAS Catalog of Selected Documents in Psychology, 7* (No. 1439).

Hoppock, R. (1935). *Job satisfaction*. New York, NY: Harper & Row.

Jones, R. A., Sensenig, J., & Ashmore, R. D. (1978). Systems of values and their multidimensional representations. *Multivariate Behavioral Research, 13*, 255–270. doi:10.1207/s15327906mbr1303_1

Kahle, L. R. (1983). *Social values and social change: Adaptation to life in America*. New York, NY: Praeger Publishers.

Kendall, M. G., & Babington Smith, B. (1940). On the method of paired comparisons. *Biometrika, 31*, 324–345. doi:10.2307/2332613

Kitwood, T. M., & Smithers, A. G. (1975). Measurement of human values: An appraisal of the work of Milton Rokeach. *Educational Research, 17*, 175–179. doi:10.1080/0013188750170302

Kluckhorn, C. K. M. (1951). Values and value orientation in the theory of action. In T. Parsons & R. Shils (Eds.), *Toward a general theory of action* (pp. 388–433). Cambridge, MA: Harvard University Press.

Knowdell, R. L. (1998). *Career Values Card Sort planning kit.* San Jose, CA: Career Research and Testing.

Kopelman, R. E., Rovenpor, J. L., & Guan, M. (2003). The study of values: Construction of the fourth edition. *Journal of Vocational Behavior, 62*, 203–220. doi:10.1016/S0001-8791(02)00047-7

Kraut, A. I., & Ronen, S. (1975). Validity of job facet importance: A multinational, multicriteria study. *Journal of Applied Psychology, 60*, 671–677. doi:10.1037/0021-9010.60.6.671

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlations hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*, 526–538. doi:10.1086/269128

Lee, J. A., Soutar, G., & Louviere, J. (2008). The best-worst scaling approach: An alternative to Schwartz's Values Survey. *Journal of Personality Assessment, 90*, 335–347. doi:10.1080/00223890802107925

Lee, J. A., Soutar, G. N., & Louviere, J. (2007). Measuring values using best-worst scaling: The LOV example. *Psychology and Marketing, 24*, 1043–1058. doi:10.1002/mar.20197

Leuty, M. E., & Hansen, J. C. (2011). Evidence of construct validity for work values. *Journal of Vocational Behavior, 79*, 379–390. doi:10.1016/j.jvb.2011.04.008

Lindeman, M., & Verkasalo, M. (2005). Measuring values with the short Schwartz's Value Survey. *Journal of Personality Assessment, 85*, 170–178. doi:10.1207/s15327752jpa8502_09

Lofquist, L. H., & Dawis, R. V. (1978). Values as second-order needs in the Theory of Work Adjustment. *Journal of Vocational Behavior, 12*, 12–19. doi:10.1016/0001-8791(78)90003-9

Louviere, J. J., & Woodworth, G. G. (1990). *Best-worst scaling: A model for largest difference judgments* (Working Paper). University of Alberta, Edmonton, Alberta, Canada.

Macnab, D., & Fitzsimmons, G. W. (1987). A multitrait-multimethod study of work related needs values and preferences. *Journal of Vocational Behavior, 30*, 1–15. doi:10.1016/0001-8791(87)90022-4

Maio, G. R., Roese, N. J., Seligman, C., & Katz, A. (1996). Rankings, ratings, and the measurement of values: Evidence for the superior validity of ratings. *Basic and Applied Social Psychology, 18*, 171–181. doi:10.1207/s15324834basp1802_4

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology, 49*, 464–480. doi:10.1016/j.jmp.2005.05.003

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*, 370–396. doi:10.1037/h0054346

Maslow, A. H. (1954). *Motivation and personality*. New York, NY: Harper & Row.

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly*, *64*, 271–298. doi:10.1086/317989

McCloy, R., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999a). *Development of the O*NET Paper-and-Pencil Work Importance Locator.* Raleigh, NC: National Center for O*NET Development. Retrieved from http://www.onetcenter.org/dl_files/DevWIL.pdf

McCloy, R., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999b). *Development of the O*NET computerized Work Importance Profiler.* Raleigh, NC: National Center for O*NET Development. Retrieved from http://www.onetcenter.org/dl_files/DevCWIP.pdf

McIntyre, S. H., & Ryans, A. B. (1977). Time and accuracy measures of alternative multidimensional scaling data collection methods: Some additional results. *Journal of Marketing Research, 14*, 607–610. doi:10.2307/3151210

Miethe, T. D. (1985). The validity and reliability of value measurements. *Journal of Psychology: Interdisciplinary and Applied, 119*, 441–453.

Miller, W. R., C'de Baca, J., Matthews, D. B., & Wilbourne, P. L. (2001). *Personal Values Card Sort.* Retrieved from http://casaa.unm.edu

Miner, J. B. (2006). *Organizational behavior 1: Essential theories of motivation and leadership.* New York, NY: Sharpe.

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the management of personal values in cross-cultural marketing. *Journal of Marketing Research, 16*, 48–52. doi:10.2307/3150873

Murray, H. A. (1938). *Explorations in personality*. New York, NY: Oxford University Press.

Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology, 7*, 403–414. doi:10.1080/1364557032000081654

Rankin, W. L., & Grube, J. W. (1980). A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology, 10*, 233–246. doi:10.1002/ejsp.2420100303

Robinson, C. H., & Betz, N. E. (2008). A psychometric evaluation of Super's Work Values

Inventory—Revised. *Journal of Career Assessment, 16*, 456–473. doi:10.1177/1069072708318903

Roe, R. A., & Ester, R. (1999). Values at work: Empirical findings and theoretical perspective. *Applied Psychology, 48*, 1–21. doi:10.1111/j.1464-0597.1999.tb00046.x

Rokeach, M. (1973). *The nature of human values.* New York, NY: Free Press.

Rokeach, M. (1982). On the validity of Spranger-based measures of value similarity. *Journal of Personality and Social Psychology, 42*, 88–89. doi:10.1037/0022-3514.42.1.88

Ronen, S., Kraut, A. I., Lingoes, J. C., & Aranya, N. (1979). A nonmetric scaling approach to taxonomies of employee work motivation. *Multivariate Behavioral Research, 14*, 387–401. doi:10.1207/s15327906mbr1404_1

Rounds, J. B. (1990). The comparative and combined utility of work value and interest data in career counseling with adults. *Journal of Vocational Behavior, 37*, 32–45. doi:10.1016/0001-8791(90)90005-M

Rounds, J. B., & Armstrong, P. I. (2005). Assessment of needs and values. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling* (pp. 305–329). Hoboken, NJ: Wiley.

Rounds, J. B., Henley, G. A., Dawis, R. V., Lofquist, L. H., & Weiss, D. J. (1981). *Manual for the Minnesota Importance Questionnaire: A measure of vocational needs and values.* Minneapolis: Department of Psychology, University of Minnesota.

Schaffer, R. H. (1953). Job satisfaction as related to need satisfaction in work. *Psychological Monographs: General and Applied, 67*, 1–29. doi:10.1037/h0093658

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–65). New York, NY: Academic Press.

Schwartz, S. H. (1994). Beyond individualism–collectivism: New cultural dimensions of values. In U. Kim, H. C. Triandis, C. Kagitcibasi, S. C. Choi, & G. Yoon (Eds.), *Individualism and collectivism: Theory, method and applications* (pp. 85–119). London, England: Sage.

Schwartz, S. H., & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology, 32*, 268–290. doi:10.1177/0022022101032003002

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38*, 230–255. doi:10.1016/S0092-6566(03)00069-2

Schwartz, S. H., Verkasalo, M., Antonovsky, A., & Sagiv, L. (1997). Value priorities and social desirability: Much substance, some style. *British Journal of Social Psychology, 36*, 3–18. doi:10.1111/j.2044-8309.1997.tb01115.x

Seaburg, D. J., Rounds, J. B., Jr., Dawis, R. V., & Lofquist, L. H. (1976, September). *Values as second order needs.* Paper presented at the 84th Annual Convention of the American Psychological Association, Washington DC.

Spini, D. (2003). Measurement equivalence of 10 value types from the Schwartz Value Survey across 21 countries. *Journal of Cross-Cultural Psychology, 34*, 3–23. doi:10.1177/0022022102239152

Spranger, E. (1928). *Types of men: The psychology of ethics and personality* (P. J. W. Pigors, Trans.). Halle, Germany: Max Niemeyer Verlag.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* Chicago, IL: University of Chicago Press.

Super, D. E. (1953). A theory of vocational development. *American Psychologist, 8*, 185–190.

Super, D. E. (1957). *The psychology of careers.* New York, NY: Harper & Row.

Super, D. E. (1970). *Manual, Work Values Inventory.* Chicago, IL: Riverside.

Super, D. E. (1973). The work values inventory. In D. Zytowski (Ed.), *Contemporary approaches to interest measurement* (pp. 189–205). Minneapolis: University of Minnesota Press.

Super, D. E. (1980). A life-span, life-space, approach to career development. *Journal of Vocational Behavior, 16*, 282–298. doi:10.1016/0001-8791(80)90056-1

Super, D. E. (1985). Coming of age in Middletown: Careers in the making. *American Psychologist, 40*, 405–414. doi:10.1037/0003-066X.40.4.405

Super, D. E., Savickas, M. L., & Super, C. M. (1996). The life-span, life-space approach to careers. In D. Brown, L. Brooks, & Associates. (Eds.), *Career choice and development* (3rd ed., pp. 121–178). San Francisco, CA: Jossey-Bass.

Swanson, J. L. (1992). Vocational behavior, 1989–1991: Life-span career development and reciprocal interaction of work and nonwork. *Journal of Vocational Behavior, 41*, 101–161. doi:10.1016/0001-8791(92)90017-T

Thompson, B., Levitov, J. E., & Miederhoff, P. A. (1982). Validity of the Rokeach Value Survey. *Educational and Psychological Measurement, 42*, 899–905. doi:10.1177/001316448204200325

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. doi:10.1037/h0070288

Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*, 47–58. doi:10.1037/h0070288

U.S. Department of Labor. (2000a). *Work Importance Locator: User's guide*. Washington, DC: Employment and Training Administration. Retrieved from http://www.onetcenter.org/WIL.html?p=3

U.S. Department of Labor. (2000b). *Work Importance Profiler: User's guide*. Washington, DC: Employment and Training Administration. Retrieved from http://www.onetcenter.org/WIP.html?p=3

Vernon, P. E., & Allport, G. W. (1931). A test for personal values. *Journal of Abnormal and Social Psychology, 26*, 231–248. doi:10.1037/h0073233

Wahba, M. A., & Bridwell, L. G. (1976). Maslow reconsidered: A review of research on the need hierarchy theory. *Organizational Behavior and Human Performance, 15*, 212–240. doi:10.1016/0030-5073(76)90038-6

Weiss, D. J., Dawis, R. V., Lofquist, L. H., & England, G. W. (1966). Instrumentation for the theory of work adjustment. *Minnesota Studies in Vocational Rehabilitation* (No. 21). Retrieved from http://www.psych.umn.edu/psylabs/vpr/monograph.htm

Zytowski, D. G. (2006). *Super's Work Values Inventory—Revised user's manual.* Retrieved from http://www.kuder.com/solutions/kuderassessments.html#supers_work_values_inventory

CHAPTER 22

ASSESSMENT OF SELF-EFFICACY

Nancy E. Betz

Albert Bandura's (1977) self-efficacy theory has been one of the most significant additions to the study and understanding of individual differences in behavior and, especially, in human agency, in the past 30 years. Developed originally (Bandura, 1977) to assist in the understanding of behavioral change through cognitive–behavioral interventions, the concepts and measures of self-efficacy are now used in many areas of psychology as well as in fields such as counseling and education. The present chapter begins with a brief review of self-efficacy theory followed by a discussion of means of and issues in its measurement. It then proceeds to discuss some specific domains and measures of self-efficacy.

Briefly, as originally proposed by Bandura (1977), self-efficacy expectations refer to a person's beliefs concerning his or her ability to successfully perform a given task or behavior. These efficacy beliefs are behaviorally specific rather than general. The concept of self-efficacy must therefore have a behavioral referent to be meaningful. Perceived self-efficacy could be addressed with respect to mathematics, initiating social interactions, investing in stocks, or fixing a flat tire. Because they are discussed in reference to a specific behavioral domain, the number of different kinds of self-efficacy expectations is limited only by the possible number of behavioral domains that can be defined—in other words, it is infinite for all practical purposes.

The concept of self-efficacy expectations is particularly useful both theoretically and practically, first of all, because of its postulated behavioral consequences. These are (a) approach versus avoidance behavior; (b) quality of performance of behaviors in the target domain; and (c) persistence in the face of obstacles or disconfirming experiences (Bandura, 1977, 1997). Thus, low self-efficacy expectations regarding a behavior or behavioral domain are postulated to lead to avoidance of those behaviors, poorer performance, and a tendency to "give up" when faced with discouragement or failure. Also important is the postulated initial development of expectations of self-efficacy, by means of four sources of efficacy information: (a) performance accomplishments (i.e., experiences of successfully performing the behaviors in question), (b) vicarious learning or modeling, (c) verbal persuasion (e.g., encouragement and support from others), and (d) lower levels of emotional arousal (i.e., lower levels of anxiety) in connection with the behavior. These sources provide a framework for the design of interventions for increasing self-efficacy expectations in a given behavioral domain.

More generally, Bandura views self-efficacy as a central ingredient of personal control or agency in one's life. As Bandura (1997) stated:

> People make causal contributions to their own psychosocial functioning through mechanisms of personal agency. Among the mechanisms of agency, none is more central or pervasive than beliefs of personal efficacy. Unless people believe they can produce desired effects by their actions, they have little incentive to act. Efficacy belief, therefore, is a major basis

DOI: 10.1037/14048-022
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

> of action. People guide their lives by their beliefs of personal efficacy. *Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce a given attainment*. (pp. 2–3)

Because the theory enables us to understand and predict problem areas and to design interventions based on the sources of efficacy information, high-quality measurement of its central construct (i.e., expectations of self-efficacy with respect to a specified behavioral domain) is crucial to research and practice. However, the assessment of self-efficacy is not an easy task and is widely misunderstood and misapplied.

ISSUES IN ASSESSMENT OF SELF-EFFICACY

In psychological measurement, content and construct validity are both dependent on careful definition of the construct of interest. In the case of measures of self-efficacy, the items reflect specific behaviors from a defined behavioral domain, so measurement must begin with careful definition and delineation of the domain of interest. There is no such thing as a measure of "self-efficacy." Rather, the assessment of perceived self-efficacy derives from the researcher's interest in a specific behavioral domain. As stated by Bandura (2006), "the efficacy belief system is not a global trait but a differentiated set of self-efficacy beliefs linked to distinct realms of functioning" (p. 307). When the researcher becomes interested in a domain for which no appropriate measure of perceived self-efficacy exists, then he or she must define the domain, especially with reference to its important constituent behaviors, so that self-efficacy with reference to that domain can be assessed.

Here the principles of scale construction and content validity are important. Specifically, as described in such resources as Nunnally and Bernstein (1994) and Betz (1996), construct-oriented scale construction begins with a careful, specific yet comprehensive definition of the domain of behavior of interest; for example, mathematics or caregiving. Note that, as with any construct of interest, there is no one correct definition; rather, it is essential that the scale constructor either be a subject matter expert or collaborate with subject matter experts in formulating the definition. Behavioral items are written on the basis of the definition.

In some cases, it may be helpful to seek the input of the target individuals of the research. Flanagan's (1954) critical incident approach has been used by researchers developing measures of self-efficacy for coping with traumatic events: To develop a content valid measure, the researchers need to ask the survivors of these events to describe the kinds of challenges they faced. Questions such as, "What were the most difficult parts of surviving the hurricane?" or "What were some of the toughest challenges of surviving domestic violence?" may be asked. These descriptions can then be used to construct coping tasks that actually reflect the experience of those going through them. Other individuals whose input may be sought include the mental health professional helping the survivors—they too are in an excellent position to describe specific challenges faced. In the study of Borgogni, Pettita, and Mastrorilli (2010), the critical incident technique was used with managers to generate a list of the critical competencies for aircraft technicians in the Italian Air Force. Few psychologists would be assumed to have the expertise to specify such critical and specialized behaviors. Thus, using such individuals in task development helps to assure that the tasks reflect the critical experiences faced.

Bandura's original theoretical discussion was focused on the use of self-efficacy theory in the treatment of clinical phobias, such as snake phobia and varieties of agoraphobia (see Bandura, Adams, Hardy, & Howells, 1980). Because they were used to construct behavioral hierarchies for use in systematic desensitization, items were organized into levels of difficulty. For example, in the treatment of agoraphobics, Bandura et al. (1980) illustrated "venturing into public territory" with a graded series of tasks, including walking a few steps beyond the door of the treatment center, to the sidewalk, one fourth block, and so forth, and finally to completing a one half-mile course (p. 53). Similarly, Bandura (2006) suggested a graded series of tasks used to measure "driving self-efficacy." Such tasks might include

driving on a neighborhood street (easiest), driving on a main road, and driving on a busy interstate (most difficult). In other cases items can be ordered in terms of empirically determined difficulty (Steffen, McKibbin, Zeiss, Gallagher-Thompson, & Bandura, 2002; Turner, Betz, Edwards, & Borgen, 2010). Such ordering is particularly useful when efficacy responses are used to guide treatment programs.

After development of a new measure of self-efficacy, traditional methods of evaluation should be used, including internal consistency and test–retest reliability, and construct validity (see Cronbach & Meehl, 1955; Nunnally & Bernstein, 1994). See also Chapters 1 through 4 in this volume for detailed coverage of methods of evaluating tests and measures.

Once the behavior domain is defined and delineated, the response continuum is specified. In Bandura's (1977) original theory, level and strength of self-efficacy were distinguished. Level was assessed by a "yes" or "no" response to the question, "Can you successfully perform this behavior?"—level referred to the most difficult task the individual perceived himself or herself as able to perform in a sequence of progressively more difficult tasks. Strength referred to the individual's confidence in that perceived capability. Because of the close relationship of level and strength ratings, and because strength (Confidence) provides a continuous rather than dichotomous item response, most measures of self-efficacy used currently use a 5- to 100-point confidence continuum. Bandura (2006) recommended a 100-point confidence continuum where confidence is assessed in 10-unit intervals ranging from 0 (*cannot do at all*) to 100 (*highly certain can do*), although he also suggested that some researchers may want to collapse this to a 10-point scale. An illustrative five-level response continuum ranging from 1 (*no confidence at all*) to 5 (*complete confidence*) is used in Paulsen and Betz's (2004) study of career decision-making self-efficacy. Bandura (2006) recommended beginning with a practice item where participants provide a 0- to 100-point rating in their confidence that they can accomplish each of a graded series of tasks, such as lifting a 10-lb. object, lifting a 20-lb. object, and so forth (p. 320).

There has been some research regarding the psychometric quality of shorter versus longer response continua. Pajares, Hartley, and Valiente (2001) compared the 0–100-point scale and a 6-point Likert-type scale in measuring writing self-efficacy in middle school students. Although the two response scales were equally reliable and had the same factor structure, the 0–100 scale was more strongly related to achievement indices, and it was related to achievement in a regression model, whereas the 6-point scale was not.

In contrast, Betz, Hammond, and Multon (2005) compared five-level and 10-level response continua to the 25-item (five subscales) short form of the Career Decision Self-Efficacy Scale (Betz, Klein, & Taylor 1996). A total of five samples, nearly 2,300 college students from several campuses, was used. Values of coefficient alpha values across the subscales ranged from.78 to .87 using the 5-point continuum and .69 to .83 for the 10-point continuum. Correlations with career indecision and clarity of vocational identity were comparable for the two response continua. Pajares and Miller (1995) used a 5-point versus a 10-point response continuum in their revision of the Math Self-Efficacy Scale (Betz & Hackett, 1983) and found no loss in internal consistency reliability. It may be that a comparison between 5- and 10-point continua provides less contrast effect than that between the 100-point and 6-point scales used by Parajes et al. (2001). In any case, it may be noted that in the research covered in this review, the large majority used Likert-type scales with 5 to 10 response points.

Following assembly of a behavioral item set, initial evaluation of item quality based on item total correlations, means and variability of obtained responses, and comprehension for the target audience should be utilized to refine the item set. Administration to a development sample should be used to evaluate internal consistency and/or test–retest reliability, and validity studies should be designed and implemented.

Most instrument development research now includes principal-components analysis or exploratory factor analysis and confirmatory factor analysis (CFA) to examine scale dimensionality (Bandura, 2006). Many studies use structural

equation models (SEMs) to examine construct validity. Thus, those constructing and evaluating measures of self-efficacy are using increasingly rigorous statistical methods.

To summarize, applications of self-efficacy theory to human behavior are theoretically infinite, yet each new application requires both the ability to carefully define the behavior domain in question and knowledge of methods of scale construction and evaluation. Because both areas of expertise are required, collaboration among researchers bringing different area of competence may be desirable.

ILLUSTRATIVE DOMAINS AND MEASURES

Self-efficacy theory has now been applied to many behavioral domains across the fields of psychology and education. To provide a flavor of these applications, self-efficacy has been measured in behavioral domains including, although not limited to, self-efficacy in stressful life transitions (Jerusalem & Mittag, 1995), self-efficacy for health-enhancing behaviors (Bandura, 2004), self-efficacy related to both the development and cessation of addictive behaviors (e.g., Schwarzer & Luszczynska, 2006), and self-efficacy for caregiving (Steffen et al., 2002). Meta-analyses have documented the utility of efficacy beliefs as predictors of effective performance and healthy functioning across a wide range of behavioral domains, including team efficacy (Gully, Incalcaterra, Joshi, & Beaubien, 2002), academic outcomes (Multon, Brown, & Lent, 1991), work-related behavior (Sadri & Robertson, 1993) and performance (Stajkovic & Luthans, 1998), career behavior (Lent, Brown, & Hackett, 1994), health outcomes (Holden, 1992), and sport performance (Moritz, Feltz, Fahrbach, & Mack, 2000).

In the next sections, the measurement of illustrative domains of self-efficacy is briefly reviewed. The length of this chapter does not permit a comprehensive review of any, much less all, of these self-efficacy domains; that would easily require several books. Rather, a sampling of measures from different areas of psychology will be provided. These areas include educational psychology, health psychology, career psychology, clinical psychology, social psychology, and gerontology. For each area, one or more illustrative measures and some representative findings are provided, but by no means is this meant to be a comprehensive review.

Self-Efficacy for Academic Achievement and Career Behavior

Self-efficacy theory has been used extensively in the study of academic achievement for elementary and secondary school children to college students. Large-scale reviews consistently support the important influence of perceived self-efficacy on motivation, learning, and achievement in education (e.g., Schunk & Pajares, 2002). Many applications within the field of education have been studied: academic functioning (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), self-efficacy for middle school math and science (Fouad, Smith, & Enochs, 1997), writing (Pajares et al., 2001), and coaching (e.g., Myers, Wolfe, & Feltz, 2005). Results uniformly support the specific effects of self-efficacy on achievement and performance (e.g., Bandura et al., 1996).

Self-efficacy regarding academic performance, particularly with respect to such domains as mathematics and technology, has also been shown to have a huge effect on college major as career choice. Domains studied within the field of career development have included mathematics self-efficacy (Lopez, Lent, Brown, & Gore, 1997), career decision self-efficacy (Luzzo, 1993), career search efficacy (Solberg, Good, Fischer, Brown, & Nord, 1995), and self-efficacy for the Holland vocational personality types (Betz, Harmon, & Borgen, 1996). Bandura, Barbaranelli, Caprara, and Pastorelli (2001) studied the relation between academic self-efficacy and middle school children's academic and career aspirations, finding that the former was related to higher aspirations for both academic and career achievements.

An illustrative scale is the Mathematics Self-Efficacy Scale (MSES; Betz & Hackett, 1983). The importance of math background to a range of educational and career options has led to its being called the "critical filter" to career development (Sells, 1982), yet many students, especially women, avoid taking math courses (Betz & Hackett, 1983). Betz and Hackett (1983) postulated that math self-efficacy

expectations have a critical role in career development.

The MSES is a three-part, 52-item measure of math self-efficacy: everyday math tasks, math courses, and math problems. Responses were obtained on a 10-point confidence scale ranging from 1 (*no confidence at all*) to 10 (*complete confidence*). Sample items on the Everyday Math Tasks subscale include balancing a checkbook and figuring out how much sales tax is owed on a purchase. The Math Courses subscale items ask the individual to provide his/her level of confidence that he/she could get an A or B in courses such as calculus and statistics. For the Math Problems subscale, confidence in solving sample math word problems is solicited. The scales have very high levels of reliability; Betz and Hackett (1983) reported coefficient alpha values of .90, .93, .and .92 for the Everyday Tasks, Math Courses, and Math Problems subscales, respectively; and Hackett and O'Halloran (1989) reported 2-week test–retest reliability coefficients of .79, .91, and .82, respectively.

Although much research was conducted on the original MSES, consistently supporting the postulates of Bandura's theory (e.g., Hackett, 1985; Lent, Brown, & Gore, 1997), Pajares and Miller (1995) developed and studied a slightly revised version of the scale (the MSES-R). Among other revisions, they used a 5-point rather than a 10-point Likert-type measure of confidence and found no loss in reliability: Coefficient alpha values were 94, .91, and .91 for the Everyday Tasks, Math Courses, and Math Problems subscales, respectively. Although the Everyday Tasks subscale was highly reliable, Pajares and Miller (1995) reported in a sample of 391 college students that the subscale was less closely related to criterion measures (performance) than were the Math Courses and Math Problems subscales. The latter two were, additionally, each most closely related to direct indices of performance. That is, the Math Courses subscale was related to choices of math-related majors, and the Math Problems subscale was most closely related to performance when the student was asked to solve those problems.

Thus, Pajares and Miller (1995) demonstrated strong support for Bandura's stress on the predictive utility of using tasks closely matched to the criterion performance of interest. As they stated: "There are different ways of assessing self-efficacy, but the most theoretically appropriate and empirically warranted is one in which the self-efficacy measure assesses the same or similar skills required for the performance task" (p. 196).

Self-Efficacy for Social Interactions

Bandura (1997) has argued persuasively for the importance of social self-efficacy in psychological adjustment, including symptoms of anxiety, phobias, and depression, and for a general sense of personal agency. He postulates an agentic model of depression in which such mechanisms as perceived self-efficacy, both social and academic, are crucial to one's sense of personal control over his or her destiny and successful adaptation to life events (Bandura, 1997). This sense of personal control and successful adaptation are, in turn, important buffers against depression.

Bandura et al. (1999) tested this model in children, using the 37-item Multidimensional Scales of Self-Efficacy (Bandura et al., 1996). The scales tap three basic domains of functioning: perceived academic self-efficacy (perceived ability to master academic subjects), perceived social self-efficacy (perceived capability for peer relationships), and perceived self-regulatory self-efficacy (perceived capability to resist pressure to engage in high-risk activities). Responses were obtained with a 5-point format from low to high belief in capability to execute the designated activities. Coefficient alpha values were .89, .82, and .70 for academic self-efficacy, social self-efficacy, and self-regulatory self-efficacy, respectively.

Testing a causal model in a sample of 282 children with a mean age of 11.5 years, Bandura et al. (1999) found that low perceived social and academic self-efficacy beliefs contributed to concurrent and later depression both directly and indirectly through their effect on prosocial behavior, academic achievement, and problem behaviors. Low social self-efficacy was a stronger predictor of depression in boys than girls. The researchers concluded that "a persistent sense of personal inefficacy operates as a common contributor to both clinical and less severe forms of depression" (Bandura et al., 1999, p. 267).

In a later longitudinal study of 650 young adolescents using the same measures, Vecchio, Gerbino, Pastorelli, Del Bove, and Caprara (2007) reported that academic and social self-efficacy were strong predictors of life satisfaction in later adolescence and were better predictors than were actual academic performance and popularity with peers.

Social self-efficacy has also been found to be important to the mental health and adjustment of adolescents and adults. For example, research has shown social self-efficacy to be related to perceived social acceptance, general self-worth, cognitive and physical competence, and self-esteem (Connolly, 1989) and, negatively, to depressive symptomatology (McFarlane, Bellissimo, & Norman, 1995).

Caprara and Steca (2005) studied the role of perceived interpersonal self-efficacy and affective self-regulatory efficacy (discussed in the next section) in the prosocial behavior and life satisfaction of four age groups ranging from young adults, adults, middle-aged adults, and elderly adults. They measured social self-efficacy with 14 items assessing perceived capability to share personal experiences with others, to invite people to go out, and to know people in a new situation. Responses were obtained on a 5-point scale ranging from 1 (*perceived incapability*) to 5 (*complete self-assurance*). The alphas across the four age groups ranged from .87 to .93. Results of path analysis indicated that social self-efficacy was related to prosocial behavior, which was, in turn, related to life satisfaction (it should be noted that affective self-regulatory efficacy was a strong predictor of both social and empathic self-efficacy, both of which influenced prosocial behavior).

Self-Regulatory Efficacy

Efficacy for self-regulation, one of the cornerstones of Bandura's (1997) theory, has received considerable research attention. As stated by Maes and Karoly (2005), "Self-regulation can be defined as a goal-guidance process aimed at the attainment and maintenance of personal goals" (p. 267). Efficacy for self-regulation has been studied in relationship to health-promoting and maintenance behaviors, affective self-regulation (Bandura, Caprara, Barbaranelli, Gerbino, & Pastorelli, 2003), and self-regulated learning (Usher & Pajares, 2008), among other domains. Evidence has consistently shown that efficacy beliefs contribute to motivation and goal-related performance though effective self-regulatory behaviors (Bandura & Locke, 2003). An overview of research on one critically important behavior domain—that of efficacy for health self-regulation—follows, in addition to mention of an illustrative scale.

As discussed by Bandura (2005), the predominance of illness in this country has shifted from acute to chronic diseases, diseases that are incredibly costly yet can, in part, be managed and controlled by effective self-regulation of health maintenance behaviors. Diseases such as diabetes, arthritis, asthma, coronary artery disease, and high cholesterol/hypertension are all syndromes that, to some or to a large extent, can be controlled through effective health care habits and avoidance of health-damaging behaviors (e.g., smoking or excessive salt intake). Bandura (2005) and Maes and Karoly (2005) summarized research directed at interventions to increase individuals' skill in and efficacy for such self-regulation.

As an example, Luszczynska and Tryburcy (2008) studied an intervention designed to increase self-efficacy for exercise and, it is hoped, the amount of exercise in which individuals engaged. They compared the intervention when used with individuals diagnosed with diabetes or cardiovascular disease with its use with those who did not have this diagnosis. A four-item scale measuring self-efficacy for exercise was constructed; subjects were asked how certain they were that they would be able to exercise under certain conditions (e.g., if their schedule was not planned for exercise or if they felt lazy). The response scale ranged from 1 (*definitely not*) to 4 (*exactly true*). Results indicated that the intervention was effective for those with the diagnoses but not for the others. Also, the effects of the intervention on amount of exercise were mediated by increases in self-efficacy. Coefficient alpha values were .88 at both pretest and posttest.

Coping Self-Efficacy

One important application of self-efficacy theory has been the study of the role of coping self-efficacy in psychological responses to trauma. The kinds of trauma investigated have included natural disasters

(e.g., hurricanes), terrorist attacks, interpersonal violence and abuse, and military combat. As defined by Benight, Swift, Sanger, Smith, and Zeppelin (1999), coping self efficacy is "the perception of one's capability for managing stressful or threatening environmental demands" (p. 2444); Benight and Bandura (2004) have reviewed evidence from many studies demonstrating the generalized role of coping self-efficacy in reduced psychological distress and improved adaptability of responses to such stresses. As is generally true in self-efficacy measurement, instruments must be tailored to tap the specific demands and stressors associated with a given type of trauma.

Benight, Ironson, and Durham (1999) developed a measure of self-efficacy for coping with the demands faced by hurricane survivors. They developed an item set by interviewing panels of psychology and psychiatry professors, graduate students on the project, and hurricane victims. Items were based on the following instructions: "For the following situations rate how confident you are that you can successfully deal with them" (p. 381); response options ranged from 1 (*not at all capable*) to 7 (*totally capable*). Situational demands included ensuring personal safety and finding shelter and food. Respondents were 288 survivors of Hurricanes Opal and Andrew. Both samples were subject to principal-components analysis, with one factor resulting in both cases. The reliability (coefficient alpha) of the seven-item scale was .87. Scores were positively related to optimism and social support and negatively related to psychological and trauma-related distress and pessimism.

Benight, Harding-Taylor, Midboe, and Durham (2004) measured self-efficacy for coping with domestic violence. Defining the variable as efficacy for coping with assault recovery demands, the authors began with focus groups of assault survivors and domestic violence advocates to determine the specific stresses facing an assault survivor. After factor analyzing responses to 50 initial items, a 30-item scale, the Domestic Violence Coping Self-Efficacy Scale, was developed. Stressors (items) included managing feelings of anxiety; finding shelter, food, and medical assistance after the attack; and being strong for others. Responses to items were made on a 100-point scale ranging from 0 (*not at all capable*) to 100 (*totally capable*). In a sample of 283 survivors of domestic violence, a coefficient alpha of .97 was obtained. Scores were significantly positively correlated with optimism, active coping, and healthy psychological functioning and were negatively correlated with distress, negative mood, and giving up.

Most recently, in a large-scale review of these studies utilizing a total of 8,011 participants, Luszczynska, Benight, and Cieslak (2009) reported medium to large effects of self-efficacy in reducing the severity and frequency of symptoms of posttraumatic stress disorder. Higher self-efficacy also led to better somatic health (e.g., reduced pain and fatigue) in the long term.

Caregiving Self-Efficacy

As the prevalence of chronic disease has increased, so has the need increased for caregiving by others, particularly family and friends. Research has demonstrated the toll that continuous caregiving has on both the physical and psychological health of the caregiver (Schulz, O'Brien, Bookwala, & Fleissner, 1995), and some researchers have examined whether self-efficacy for caregiving may help to mediate its ill effects.

For example, Steffen and her colleagues developing the Revised Scale for Caregiving Self-Efficacy (Steffen et al., 2002). Steffen et al. used two samples of caregivers of cognitively impaired older adults (mostly patients with Alzheimer's disease)—a total of 314 individuals—to revise and evaluate the scale originally developed by Zeiss, Gallagher-Thompson, Lovett, Rose, and McKibbin (1999). The scale measures three domains: obtaining respite (e.g., asking for help when you need it), responding to disruptive patient behaviors (e.g., remaining calm when the patient is engaging in repetitive behaviors), and controlling upsetting thoughts (e.g., controlling thoughts about a patient's lack of personal hygiene).

Scale development was exemplary. From the earlier version of the scale (Zeiss et al., 1999), Steffen et al. increased the size of the item pool and then subjected it to analysis in the first sample. Traditional item analyses, factor analysis, and examination of item response distributions (including means, standard deviations, skewness, and kurtosis) were

used in item selection. With the second sample, the following were examined using SEM: CFA; additional psychometric analysis; and convergent, discriminant, and construct validity.

Responses are obtained on a confidence scale ranging from 0 (*cannot do at all*) to 100 (*certainly can do*). Steffen and colleagues have used the scale to examine the effectiveness of self-efficacy interventions for the management of anger and depression in family caregivers of patients with dementia (e.g., Gilliam & Steffen, 2006).

Collective Efficacy

Although much organizational research has focused on self-efficacy for a given individual, there is increasing interest as well in generalizing self-efficacy to group levels of analysis. Bandura (1997) defined "collective efficacy" as "a group's shared belief in its conjoint capabilities to organize and execute the courses of action required to produce given levels of attainments" (p. 477). Collective efficacy has consistently been shown to be related to group motivation and effective performance (Bandura, 2000). It has been studied at many levels of analysis, from small work teams (Gully et al., 2002) to organized sports teams (Magyar, Feltz, & Simpson, 2004), military forces (Borgogni, Pettita, & Mastrorilli, 2010), educational institutions (Caprara, Barbaranelli, Borgogni, & Steca, 2003), and governmental organizations (Borgogni, Russo, Pettita, & Latham, 2009).

One use of the overall concept of collective efficacy is that of "team efficacy." Collective efficacy can refer to the collective beliefs systems of teams, departments, or entire organizations, whereas team efficacy refers to beliefs within a defined work team. Furthermore, team efficacy, like collective efficacy, refers to more than just the sum of the self-efficacy beliefs of the team members; rather, it describes a shared belief in the capabilities of the collective operating as a unit. Research on team efficacy suggests that it, like collective efficacy more generally, is positively related to group performance. Gully et al. (2002) undertook a meta-analysis of 76 empirical studies (256 effect sizes) of the relationship of task-specific team efficacy to performance; studies were taken from journals in industrial/organizational psychology and from management and personnel journals. Team efficacy was measured in many different ways across the 67 studies, so it is difficult to evaluate the quality of measures utilized. The overall effect size of the relation between team efficacy and performance was .41, a moderately sized effect.

Caprara et al. (2003) studied the roles of individual and collective self-efficacy in the job satisfaction of 2,688 teachers in 107 Italian junior high schools. Twelve items measured teachers' confidence in their abilities to handle various challenges in teaching, such as dealing with difficult students. To measure collective efficacy, nine items related to such objectives and supporting important initiatives in the community were used. Responses were obtained on a 7-point Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The data were analyzed using CFA and SEM. Values of coefficient alpha were .74 and .82 overall for the individual and collective self-efficacy measures, respectively, and were .92 and .95, respectively, when analyzed between, rather than within, the 103 schools. Both individual self-efficacy and collective efficacy were related to teachers' job satisfaction.

At a larger level of analysis, Borgogni et al. (2010) investigated the relationship between self-efficacy and collective efficacy to job satisfaction and commitment among 387 technicians and staff in the Italian Air Force. Like the studies of coping self-efficacy, Flanagan's (1954) critical incident technique was used in focused interviews with managers to determine the critical tasks both for individuals and the work group as a whole. Self-efficacy was measured by 17 items assessing beliefs about being able to handle job responsibilities, technical challenges, emergencies, and interpersonal relationships. A 7-point Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) was used. Collective self-efficacy referred to the individual's belief that the organization as a whole could effectively cope with problems and issues. Coefficient alphas were high: .93 for self-efficacy and .89 for collective self-efficacy. In terms of results, self-efficacy was related to collective efficacy which was related to organizational commitment and job satisfaction.

Borgogni et al. (2009) also studied 170 employees of a city hall in Italy. They defined facets of collective efficacy for the group and for the organization. They used the Flanagan (1954) critical incident technique

in focus groups asked to identify critical challenges and emergencies faced by staff and officials. For the group and organization collective efficacy scales, items dealing with effective response to emergencies and providing citizens with high-quality services were used (19 items for group efficacy and 21 for organization efficacy). The alpha for group self-efficacy was .97 and that for organizational self-efficacy was .98. Both measures of self-efficacy were significantly related to organizational commitment.

Generalized Self-Efficacy (GSE)

Ever since the origination of self-efficacy theory, there has been an urge to generalize the concept; hence, considerable research on GSE, has been done. As noted by many researchers (e.g, Chen, Gully, & Eden, 2004), although the GSE concept originated from Bandura's theory, GSE is actually a different concept than self-efficacy as conceived by Bandura, who viewed it as behaviorally specific. As noted by Chen et al. (2004), "The GSE construct originated from the concept of self-efficacy generally . . . however it is distinguishable from the concept of self-efficacy because whereas self-efficacy is a relatively malleable, task specific belief, GSE is a relatively stable, trait-like, generalized competence belief" (p. 376). A number of measures of GSE have been developed, including Sherer et al.'s (1982) General Self-Efficacy Scale; Chen et al.'s (2001) new general Self-Efficacy Scale; and Schwarzer, Baßler, Kwiatek, Schröder, and Zhang's (1997) General Self-Efficacy Scale. Chen et al. (2004) compared the predictive efficacy of global self-esteem and GSE, finding that GSE is related to motivational variables such as conscientiousness, need achievement, and performance, whereas global self-esteem is related significantly to affective variables such as positive affect and, negatively, to negative affect and anxiety. Thus, GSE may have a general usefulness in such predictions, but task specificity is likely to have greater predictive utility for specific performance criteria.

USING ITEM RESPONSE THEORY AND ADAPTIVE TESTING

This chapter has used classical test theory as the basis for evaluating reliability of the scales discussed herein, but item response theory (IRT; see Wainer, 2000; Volume 1, Chapter 6, this handbook) has important uses that have only begun to be studied in the context of the assessment of self-efficacy. In brief, IRT enables the description of items and individual in the same metric, usually that of the standard normal distribution. Item parameters of "difficulty" and discrimination" are assigned to each item: The *b* parameter is the estimate of the level of the latent trait (θ) and is described using the same metric as that of the item. IRT also provides an index of the item's discriminatory power (*a*) in the item information function (IIF) for each value of the underlying trait (θ). The sum of the IIFs of the items administered is the total test information function, which yields an index of the precision of the measurement at each point on the trait continuum. Information is the square root of the inverse of the standard error of measurement, which is inversely related to traditional internal consistency reliability coefficients, such as coefficient alpha. Thus, both item and test quality can be precisely described for individuals with different levels of the underlying trait.

One particular advantage of using IRT to evaluate these items is that adaptive testing can then be utilized (e.g., Wainer, 2000). In adaptive testing, item difficulty (also in the metric of the trait estimate) is "adapted" to the individual's response pattern: The next item administered is that closest to the individual's estimated trait level. The measurement of self-efficacy is uniquely suited to adaptive testing because self-efficacy items, like ability and aptitude test items and unlike personality and attitude test items, are easily scaled for item difficulty (e.g., see Steffen et al., 2002). As mentioned earlier, Bandura's (1977) original discussions of self-efficacy theory were in relationship to the treatment of clinical phobias through systematic desensitization. This method involves the development of a "fear hierarchy" by the patient; that is an organized series of tasks for which desensitization or cognitive–behavioral intervention proceeds from the easiest to the most difficult tasks.

Conceptually, therefore, self-efficacy theory assumes that tasks can be organized in a "difficulty" hierarchy for any given individual or group. The overall confidence judgments elicited by behavioral

items are the index of item difficulty. Adaptive testing provides an excellent alternative to the administration of long scales, which can be unnecessarily costly. Turner et al. (2010) simulated the use of adaptive testing with an inventory of career self-efficacy, the Career Confidence Inventory (Betz & Borgen, 2010), a 190-item inventory measuring self-efficacy or confidence with respect to the six Holland vocational personality types: realistic (mechanical, outdoors), investigative (math and science), artistic (music, art, drama, writing), social (helping, socializing), enterprising (persuading, management, entrepreneurship), and conventional (detail work with both verbal and mathematical content). The entire inventory takes 20 to 25 minutes to administer. Using adaptive testing, overall reliability estimates for each scale in the .80s could be obtained with as few as 28 total items for all six types, taking about 8 minutes. When time is limited or other inventories need to be administered, this capability can be extremely valuable.

SUMMARY

Overall, this chapter illustrates not only the substantive breadth of the assessment of self-efficacy but also the high quality of those measures included herein. There were many measures of self-efficacy that were not of high quality or were described in insufficient detail to make an evaluation of quality, so they were not included in this chapter. It is hoped that researchers interested in assessment of self-efficacy will carefully study the recommendations regarding ensuring high quality in the assessment of self-efficacy, found in the first part of this chapter. Researchers who have already developed high-quality measures may consider using IRT and adaptive testing as well. This is a growing and important area of research for many areas of human functioning, and I hope that this chapter stimulates further high-quality measurement of expectations of self-efficacy.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. doi:10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: Freeman.

Bandura, A. (2000). Exercise of human agency through collective efficacy. *Current Directions in Psychological Science, 9*, 75–78. doi:10.1111/1467-8721.00064

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education and Behavior, 31*, 143–164. doi:10.1177/1090198104263660

Bandura, A. (2005). The primacy of self-regulation in health promotion. *Applied Psychology, 54*, 245–254. doi:10.1111/j.1464-0597.2005.00208.x

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 337–338). Greenwich, CT: Information Age.

Bandura, A., Adams, N., Hardy, A., & Howells, G. (1980). Tests of the generality of self-efficacy theory. *Cognitive Therapy and Research, 4*, 39–66. doi:10.1007/BF01173354

Bandura, A., Barbaranelli, C., Caprara, G., & Pastorelli, C. (1996). Multifaceted impact of self efficacy beliefs on academic functioning. *Child Development, 67*, 1206–1222. doi:10.2307/1131888

Bandura, A., Barbaranelli, C., Caprara, G., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development, 72*, 187–206. doi:10.1111/1467-8624.00273

Bandura, A., Caprara, G., Barbaranelli, C., Gerbino, M., & Pastorelli, C. (2003). Impact of affective self-regulatory efficacy on diverse spheres of functioning. *Child Development, 74*, 769–782. doi:10.1111/1467-8624.00567

Bandura, A., & Locke, E. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology, 88*, 87–99. doi:10.1037/0021-9010.88.1.87

Bandura, A., Pastorelli, C., Barbaranelli, C., & Caprara, G. V. (1999). Self-efficacy pathways to childhood depression. *Journal of Personality and Social Psychology, 76*, 258–269. doi:10.1037/0022-3514.76.2.258

Benight, C. C., & Bandura, A. (2004). Social cognitive theory of posttraumatic recovery: The role of perceived self-efficacy. *Behaviour Research and Therapy, 42*, 1129–1148. doi:10.1016/j.brat.2003.08.008

Benight, C. C., Harding-Taylor, A., Midboe, A., & Durham, R. (2004). Development and psychometric evaluation of a domestic violence coping self-efficacy measure. *Journal of Traumatic Stress, 17*, 505–508. doi:10.1007/s10960-004-5799-3

Benight, C. C., Ironson, G., & Durham, R. (1999). Psychometric properties of a hurricane coping measure. *Journal of Traumatic Stress, 12*, 379–386. doi:10.1023/A:1024792913301

Benight, C., Swift, E., Sanger, J., Smith, A., & Zeppelin, D. (1999). Coping self-efficacy as a mediator of distress following a natural disaster. *Journal*

of Applied Social Psychology, 29, 2443–2464. doi:10.1111/j.1559-1816.1999.tb00120.x

Betz, N. E. (1996). Test Construction. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (2nd ed., pp. 239–250). Thousand Oaks, CA: Sage Publications, Inc.

Betz, N. E., & Borgen, F. (2010). The CAPA integrative online system for college major exploration. *Journal of Career Assessment*, *18,* 317–327. doi:10.1177/1069072710374492

Betz, N. E., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior, 23*, 329–345. doi:10.1016/0001-8791(83)90046-5

Betz, N. E., Hammond, M., & Multon, K. (2005). Reliability and validity of response continua for the Career Decision Self-Efficacy Scale. *Journal of Career Assessment, 13*, 131–149. doi:10.1177/1069072704273123

Betz, N. E., Harmon, L., & Borgen, F. (1996). The relationships of self-efficacy for the Holland themes to gender, occupational group membership, and vocational interests. *Journal of Counseling Psychology, 43*, 90–98. doi:10.1037/0022-0167.43.1.90

Betz, N. E., Klein, K., & Taylor, K. (1996). Evaluation of a short form of the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment, 4*, 47–57. doi:10.1177/106907279600400103

Borgogni, L., Pettita, L., & Mastrorilli, A. (2010). Correlates of collective efficacy in the Italian Air Force. *Applied Psychology, 59*, 515–537.

Borgogni, L., Russo, S., Pettita, L., & Latham, G. (2009). Collective efficacy and organizational commitment in an Italian City hall. *European Psychologist, 14*, 363–371. doi:10.1027/1016-9040.14.4.363

Caprara, G., & Steca, P. (2005). Self-efficacy beliefs as determinants of prosocial behavior conducive to life satisfaction across ages. *Journal of Social and Clinical Psychology, 24*, 191–217. doi:10.1521/jscp.24.2.191.62271

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Steca, P. (2003). Efficacy beliefs as determinants of teachers' job satisfaction. *Journal of Educational Psychology, 95*, 821–832. doi:10.1037/0022-0663.95.4.821

Chen, G., Gully, S., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4*, 62–83. doi:10.1177/109442810141004

Chen, G., Gully, S., & Eden, D. (2004). Generalized self-efficacy and self-esteem: Toward theoretical and empirical distinction between correlated self-evaluations. *Journal of Organizational Behavior, 25*, 375–395. doi:10.1002/job.251

Connolly, J. (1989). Social self-efficacy in adolescence. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 21*, 258–269. doi:10.1037/h0079809

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Flanagan, J. C. (1954). The critical incidents technique. *Psychological Bulletin*, *51*, 327–358. doi:10.1037/h0061470

Fouad, N. A., Smith, P. L., & Enochs, L. (1997). Reliability and validity evidence for the Middle School Self-Efficacy Scale. *Measurement and Evaluation in Counseling and Development, 30*, 17–31.

Gilliam, C. M., & Steffen, A. (2006). The relationship between caregiving self-efficacy and depressive symptoms in dementia family caregivers. *Aging and Mental Health, 10*, 79–86. doi:10.1080/13607860500310658

Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology, 87*, 819–832. doi:10.1037/0021-9010.87.5.819

Hackett, G. (1985). Role of mathematics self-efficacy in the choice of math-related majors of college men and women. *Journal of Counseling Psychology, 32*, 47–56. doi:10.1037/0022-0167.32.1.47

Hackett, G., & O'Halloran, M. S. (1989). The relationship of role model influences to the career Salience and educational and career plans of college women. *Journal of Vocational Behavior, 35*, 164–180. doi:10.1016/0001-8791(89)90038-9

Holden, G. (1992). The relationship of self-efficacy appraisals to subsequent health related outcomes: A meta-analysis. *Social Work in Health Care*, *16*, 53–93. doi:10.1300/J010v16n01_05

Jerusalem, M., & Mittag, W. (1995). Self-efficacy in stressful life transitions. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 177–201). Cambridge, England: Cambridge University Press.

Lent, R. W., Brown, S. D., & Gore, P. (1997). Discriminant and predictive validity of academic self-concept, self-efficacy and math-specific self-efficacy. *Journal of Counseling Psychology, 44*, 307–315. doi:10.1037/0022-0167.44.3.307

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79–122. doi:10.1006/jvbe.1994.1027

Lopez, F. G., Lent, R. W., Brown, S. D., & Gore, P. A., Jr. (1997). Role of social-cognitive expectations in high

school students' mathematics related interest and performance. *Journal of Counseling Psychology, 44*, 44–52. doi:10.1037/0022-0167.44.1.44

Luszczynska, A., Benight, C., & Cieslak, R. (2009). Self-efficacy and health related outcomes of collective trauma. A systematic review. *European Psychologist, 14*, 51–62. doi:10.1027/1016-9040.14.1.51

Luszczynska, A., & Tryburcy, M. (2008). Effects of a self-efficacy intervention on exercise: The moderating role of diabetes and cardiovascular diseases. *Applied Psychology, 57*, 644–659.

Luzzo, D. (1993). Value of career decision making self-efficacy in predicting career decision making attitudes and skills. *Journal of Counseling Psychology, 40*, 194–199. doi:10.1037/0022-0167.40.2.194

Maes, S., & Karoly, P. (2005). Self-regulation assessment and intervention in physical health and illness: A review. *Applied Psychology, 54*, 267–299. doi:10.1111/j.1464-0597.2005.00210.x

Magyar, T. M., Feltz, D. L., & Simpson, I. P. (2004). Individual and crew level determinants of collective efficacy in rowing. *Journal of Sport and Exercise Psychology*, *26*, 136–153.

McFarlane, A. H., Bellisimo, A., & Norman, G. R. (1995). The role of family and peers in social self-efficacy: Links to depression in adolescence. *American Journal of Orthopsychiatry, 65*, 402–410. doi:10.1037/h0079655

Moritz, S. E., Feltz, D. L., Fahrbach, K. R., & Mack, D. E. (2000). The relation of self-efficacy measures to sport performance: A meta-analytic review. *Research Quarterly for Exercise and Sport*, *71*, 280–294.

Multon, K., Brown, S., & Lent, R. (1991). Relation of self-efficacy beliefs to academic outcomes. *Journal of Counseling Psychology, 38*, 30–38. doi:10.1037/0022-0167.38.1.30

Myers, N. D., Wolfe, E., & Feltz, D. (2005). Evaluation of the psychometric properties of the Coaching Self-Efficacy Scale for coaches from the United States of America. *Measurement in Physical Education and Exercise Science, 9*, 135–160.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.

Pajares, F., Hartley, J., & Valiente, G. (2001). Response format in writing self-efficacy assessment. *Measurement and Evaluation in Counseling and Development, 33*, 214–221.

Pajares, F., & Miller, M. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology, 42*, 190–198. doi:10.1037/0022-0167.42.2.190

Paulsen, A. M., & Betz, N. (2004). Basic confidence predictors of career decision self-efficacy. *Career Development Quarterly, 52*, 353–361. doi:10.1002/j.2161-0045.2004.tb00951.x

Sadri, G., & Robertson, I. T. (1993). Self-efficacy and work-related behavior: A review and meta-analysis. *Applied Psychology, 42*, 139–152. doi:10.1111/j.1464-0597.1993.tb00728.x

Schulz, R., O'Brien, A., Bookwala, J., & Fleissner, K. (1995). Psychiatric and physical morbidity effects of dementia caregiving: Prevalence, correlates, and causes. *The Gerontologist, 35*, 771–791. doi:10.1093/geront/35.6.771

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 15–31). San Diego, CA: Academic Press. doi:10.1016/B978-012750053-9/50003-6

Schwarzer, R., Baßler, J., Kwiatek, P., Schröder, K., & Zhang, J. X. (1997). The assessment of optimistic self-beliefs: Comparison of the German, Spanish, and Chinese versions of the General Self-Efficacy Scale. *Applied Psychology, 46*, 69–88. doi:10.1111/j.1464-0597.1997.tb01096.x

Schwarzer, R., & Luszczynska, A. (2005). Self-efficacy, adolescents' risk-taking behaviors, and health. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 139–160). Greenwich, CT: Information Age.

Sells, L. (1982). Leverage for equal opportunity through mastery of mathematics. In S. M. Humphries (Ed.), *Women and minorities in science* (pp. 7–26). Boulder, CO: Westview Press.

Sherer, M., Maddux, J. E., Mercadante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The Self-Efficacy Scale: Construction and validation. *Psychological Reports, 51*, 663–671. doi:10.2466/pr0.1982.51.2.663

Solberg, V. S., Good, G. E., Fischer, A. R., Brown, S. E., & Nord, D. (1995). Relative effects of career search self-efficacy and human agency upon career development. *Journal of Counseling Psychology, 42*, 448–455. doi:10.1037/0022-0167.42.4.448

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*, 240–261. doi:10.1037/0033-2909.124.2.240

Steffen, A. M., McKibbin, C., Zeiss, A. M., Gallagher-Thompson, D., & Bandura, A. (2002). The Revised Scale for Caregiving Self-Efficacy: Reliability and validity studies. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *57*, 74–86. doi:10.1093/geronb/57.1.P74

Turner, B. T., Betz, N. E., Edwards, M. E., & Borgen, F. (2010). Psychometric examination of an inventory of self-efficacy for the Holland themes using item

response theory. *Measurement and Evaluation in Counseling and Development, 43,* 188–198.

Usher, E. L., & Pajares, F. (2008). Self-efficacy for self-regulated learning. *Educational and Psychological Measurement, 68*, 443–463. doi:10.1177/0013164407308475

Vecchio, G., Gerbino, M., Pastorelli, C., Del Bove, G., & Caprara, G. (2007). Multifaceted self-efficacy beliefs as predictors of life satisfaction in late adolescence. *Personality and Individual Differences, 43*, 1807–1818. doi:10.1016/j.paid.2007.05.018

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Erlbaum.

Zeiss, A. M., Gallagher-Thompson, D., Lovett, S., Rose, J., & McKibbin, C. (1999). Self-efficacy as a mediator of caregiver coping: Development and testing of an assessment model. *Journal of Clinical Geropsychology, 5*, 221–230. doi:10.1023/A:1022955817074

CHAPTER 23

ASSESSMENT OF ETHNIC IDENTITY AND ACCULTURATION

Moin Syed

Cultural identity is an important aspect of the ethnic minority and immigrant experience. After decades of research that primarily used demographic categories, such as ethnic labels (Chinese American) and immigrant generational status (born in the United States) as markers of cultural identities, current practices aim to assess the cognitive, affective, and behavioral components of how people identify with their cultural background. This chapter focuses on the two major operational definitions of cultural identity: ethnic identity and acculturation. Although similar, and often used interchangeably, ethnic identity and acculturation are distinct elements of cultural identity. *Ethnic identity* is broadly defined as the degree to which individuals identify with their ethnic group. In contrast, *acculturation* is defined as the change in cultural identity as a result of navigating two distinct cultures (e.g., Chinese and U.S. cultures).

Deriving consistent definitions for the terms *ethnicity*, *race*, and *culture* seems to be an intractable problem (see Quintana, 2007). The idea that race corresponds to natural, biologically based categories of humans is not supported by research, despite widespread acceptance of the notion both in scientific communities and society at large. As described by Markus (2008), there is a strong movement to switch the focus from the perceived biological basis of race to its socially constructed nature by using the term only in the context of social structures that contribute to stratification. When considering individuals' identification with a group based on some level of shared cultural heritage, the preferred term is *ethnicity*.[1] Accordingly, throughout this chapter, the terms *ethnicity* and *ethnic identity* are primarily used, even if the original researchers used the terms *race* and *racial identity*. Readers should be aware, however, that there is currently no agreed-upon definition of these terms.

The purpose of this chapter is to summarize the existing methods of assessing ethnic identity and acculturation and to provide suggestions for future research. Both ethnic identity and acculturation research rely on myriad assessment tools, which are almost exclusively self-report using Likert-type rating scales. Rather than providing a detailed list of all available measures, the focus of this chapter is on the most widely used instruments within the two content areas. Furthermore, the focus is on the dominant theoretical models of ethnic identity and acculturation and the degree to which the instruments adequately operationalize the theory's parameters. Readers interested in detailed information (e.g., psychometrics, response scales, sample items) about a wide variety of measures are encouraged to consult existing sources on the matter: For ethnic identity measures, see various chapters in Jackson (2006) or Landis, Bennett, and Bennett (2004); for acculturation measures, see Wallace, Pomery, Latimer, Martinez, and Salovey (2010) and Zane and

[1]Of course, these two definitions do not imply independence. See Markus (2008) for more details.

I thank Linda Juang for helpful comments on an earlier version of this chapter.

DOI: 10.1037/14048-023
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

Mak (2003). Moreover, ethnic identity and acculturation do not receive equal consideration in this chapter. The reasons for this are twofold: (a) Ethnic identity research has more theoretical diversity that informs multiple instruments; and (b) in addition to identities, acculturation research encompasses values, beliefs, attitudes, and behaviors and thus ventures into territory that goes beyond the scope of this chapter.

THEORIES AND MEASURES OF ETHNIC IDENTITY

Within the ethnic identity literature, there are three broad families of measures that are situated within different theoretical frameworks: the developmental model, social–psychological models, and stage theories. This section briefly describes each of these theories and the most common measures that are used to operationalize them. It then concludes with current challenges in the field to effective assessment of ethnic identity.

Developmental Model of Ethnic Identity

Developmental models of identity are heavily influenced by Erikson (1968). For Erikson, a healthy identity is one that is characterized by both contextual integration and temporal continuity. That is, individuals must piece together their multiple identifications into a relatively coherent whole, and should maintain a sense of continuity among their past experiences, current concerns, and future prospects. Of note, Erikson viewed identity development as beginning in earnest during adolescence yet continuing to be a salient developmental task throughout the life span.

Erikson wrote a great deal about identity and described complex notions of how identities are created and revised in accordance with social and historical contexts. Seeking to bring order to this complexity, Marcia (1966) operationalized one of Erikson's critical postulations. Erikson believed that adolescents go through a period of crisis in which they must think deeply about how to integrate their multiple identifications. Because crisis implies pathology, which was not what Erikson intended, Marcia later retermed the process as *exploration* (Marcia, 1980). In creating what became known as the identity status model, Marcia suggested that youth engage in this process of exploration on the way to settling on meaningful identity commitments. Thus, the processes of identity exploration and commitment were viewed as crucial to identity development. Marcia's great contribution was to move away from a linear conception of the identity process (i.e., the process of exploration ultimately leads to commitments) and consider the two dimensions as conceptually independent. By examining the interplay between these two processes, Marcia's identity status model provided a typological approach to identity development. The interaction between the two dimensions generated four identity statuses: achieved (high exploration, high commitment), moratorium (high exploration, low commitment), foreclosed (low exploration, high commitment), and diffused (low exploration, low commitment).

In her landmark review of the ethnic identity literature, Phinney (1990) highlighted how the existing literature was fragmented, isolated, and mostly atheoretical. In creating her developmental model of ethnic identity, Phinney suggested that Marcia's identity status model could be extended to study the degree to which individuals identify with their ethnic group, or their ethnic identity. Thus, like occupational, religious, and political identities, ethnic identity could be investigated as developing through the processes of exploration and commitment.

The Multi-Group Ethnic Identity Measure (MEIM) was introduced as an instrument that could capture the developmental model of ethnic identity (Phinney, 1992). The measure included 14 items pertaining to ethnic identity and six items measuring attitudes toward other groups. Phinney (1992) maintained that the other-group attitudes scale would be useful for understanding individuals' ethnic identities but that it was not itself a measure or dimension of ethnic identity. Accordingly, the conceptual and psychometric properties of the other-group attitudes scale are not included in this review.

The original MEIM (hereinafter, the MEIM-14) consisted of 14 statements that individuals responded to on a 4-point scale, ranging from *strongly disagree* to *strongly agree*. All but two of the

items were worded positively, and after reverse-scoring as appropriate, items were averaged so that higher values corresponded to higher levels of ethnic identity. The MEIM was designed to include three subscales: Ethnic Identity achievement (five items that tapped into both exploration and commitment), Ethnic Affirmation (seven items), and Ethnic Behaviors (two items). However, factor analysis indicated the presence of a single factor. Phinney suggested that the 14 items could be averaged together to create a global measure of ethnic identity, or alternatively the three proposed subscales could be used, as they were conceptually distinct despite the one-factor solution. Indeed, the three subscales were highly intercorrelated (*r*s = .46–.79).

The first major revision of the MEIM was reported in Roberts et al. (1999). Based on factor-analytic work with ethnically diverse early adolescents, the two negatively worded items from the MEIM-14 were dropped, and some were altered, to arrive at a new 12-item measure (hereinafter, the MEIM-12). Part of the impetus for the development of the MEIM-12 was to create a measure that was better aligned with the theoretical constructs it was meant to represent, namely exploration and commitment. Accordingly, the MEIM-12 was designed to consist of two subscales corresponding to exploration (five items) and commitment (seven items), rather than the three subscales that were part of the MEIM-14. This intention was supported by a two-factor solution with very highly correlated factors (*r*s = .70–.75). As with the MEIM-14, the authors suggested that the MEIM-12 could be used as a single, global measure of ethnic identity or could be separated into its two components, Exploration and Commitment.

Phinney and Ong (2007) recently revised the MEIM again. The impetus for this revision was to improve the reliability of the Exploration subscale, to produce an even number of items for each scale, to remove all behavioral items, and to only include items that are worded in the past tense. The revision resulted in a six-item scale (the MEIM-6), with three items for each of the Exploration and Commitment subscales. Once again, a two-factor solution was supported, although the factors remained very highly correlated ($r = .74$). Although some of these revisions were improvements (even number of items, parallel tense), the Exploration subscale may not be as strong as in previous versions. Although no formal meta-analyses have been conducted, the Exploration subscale of the MEIM-12 consistently produces alpha coefficients that are much lower ($\approx$.71–.75) than for the Commitment scales ($>$.90). This differential may be because the process of exploration is inherently more multidimensional than is commitment. It is very likely that individuals could engage in some exploratory behaviors (e.g., actively learning about one's group) and not others (e.g., being active in organizations or social groups), therefore attenuating the reliability. This finding is less likely to be the case for commitment (e.g., feeling pride in one's group is likely associated with feeling a sense of belongingness).

The MEIM-6 is not sufficiently represented in the literature, and as of now, there are no published reports of its psychometric properties other than the article describing its initial development (Phinney & Ong, 2007). Whereas the MEIM-12 was a significant improvement in the instrument, at this time, this is less clearly the case for the MEIM-6. By removing the behavioral exploration items, the dimensionality of exploration, which is inherently behavioral, has been restricted.

Throughout the history of the MEIM, researchers have focused on the continuous measures of the constructs (i.e., more or less exploration, independent of more or less commitment). More recently, researchers have begun to investigate how exploration and commitment interact to produce the identity statuses described previously. The identity statuses may be particularly useful for assessment of ethnic identity in clinical and counseling settings, as they provide a momentary snapshot of the individual's developmental level. This usefulness, however, is restricted by the fact that there are no set criteria for how to classify individuals into the statuses. The currently favored approach in research settings is to use cluster-analytic methods to sort a sample into the statuses (Seaton, Scottham, & Sellers, 2006; Syed & Azmitia, 2008; Syed, Azmitia, & Phinney, 2007; Yip, Seaton, & Sellers, 2006). In these situations, statuses are determined in a relative sense within a sample rather than in an objective sense in

relation to proscribed criteria. As a result, if the means for exploration and commitment are inflated in one sample compared with another, classification in the moratorium status in the former sample may be based on the same absolute criteria as classification into the achieved status in the latter sample. Clearly, the sample-driven relativity of the identity statuses is a major limitation to their use.

A further limitation to the ethnic identity status model is that it is unclear just how many statuses adequately represent the domain. As discussed previously, Marcia's (1966, 1980) original identity status model comprised four statuses: achieved, moratorium, foreclosed, and diffused. In the first investigation into the applicability of the identity status model for ethnic identity, Phinney (1989) found clear evidence for the achieved and moratorium statuses, but could not reliably distinguish between foreclosed and diffused. Phinney suggested that perhaps these two statuses, which are both characterized by very low levels of exploration, could be combined into a single "unexamined" status and therefore the ethnic identity status model only consists of three statuses, not four (see also Phinney, 1993). Unfortunately, the existing literature has done little to resolve the debate about the appropriate number of statuses. Research with African American high school students has consistently shown four statuses (Seaton et al., 2006; Yip et al., 2006), whereas research with ethnically diverse college students (very few of whom were African American) has consistently found evidence for three (Syed & Azmitia, 2008; Syed et al., 2007). That the samples differed along a number of dimensions does not provide clear insights into the cause of the discrepancy. Three possibilities are immediately obvious: (a) that African Americans have more differentiated ethnic identities than other ethnic groups, (b) that college students/young adults have less differentiated ethnic identities than high schoolers, and (c) a combination of both. It is also worth noting that the cluster-analytic methods that have been used are not identical across studies. Furthermore, the role of expectancy biases cannot be ignored, given the subjectivity of selecting the number of cluster and interpreting their meaning.

Some researchers have attempted to unpack the construct of commitment into two distinct elements comprised therein: affirmation and resolution (Juang & Nguyen, 2010; Lee & Yoo, 2004; Umaña-Taylor, Yazedjian, & Bámaca-Gómez, 2004). They argue that individuals may have arrived at some level of clarity about what their ethnicity means to them (resolution), but that clarity is not necessarily tied to how they feel about their ethnic group (affirmation). In other words, the measurement of commitment as a sense of clarity and positive feelings may not adequately represent the structure of individuals' ethnic identities. Recent factor-analytic work on the MEIM-14 provides initial empirical support for this argument, at least structurally (Juang & Nguyen, 2010; Lee & Yoo, 2004).

The conflation between resolution and affirmation led to the development of a new measure, the Ethnic Identity Scale (EIS; Umaña-Taylor et al., 2004). The item content of the 17-item EIS is substantively very similar to the MEIM-14 but was conceptualized as a three-factor model tapping into exploration (seven items), resolution (four items), and affirmation (six items). This measure is less established and not as widely used as the MEIM, and as a result its psychometric properties have not been examined extensively. Furthermore, the Affirmation subscale is composed solely of negatively worded items, which produces a scale score that is severely skewed and restricted in range. Thus, the benefit of having affirmation separate from resolution does not seem to be fully realized with the EIS at this time.

Social–Psychological Approaches to Ethnic Identity

In contrast to the developmental model of ethnic identity, which addresses stability and change within individuals relatively irrespective of their contexts, social–psychological approaches to ethnic identity are based in social identity theory. In particular, Tajfel's (1981; Tajfel & Turner, 1986) social identity theory was proposed as an explanatory tool for why people heighten their social group identification under conditions of threat. According to social identity theory, increasing identification with a group when in a minority situation helps overcome the perceived threat and enhance self-esteem.

Thus, the social identity approach to identity emphasizes how people identify with their group in a particular situation or context, as opposed to an enduring sense of self over time and context.

Much of the earlier identity research based in social identity theory was conducted with artificially created groups in the laboratory (e.g., Tajfel, Billig, Flament, & Bundy, 1971; see also Ashmore, Deaux, & McLaughlin-Volpe, 2004, for a review). Although this procedure provided insights into how identities could be activated in situ, it failed to address directly the question of real-life social groups that occupy minority position and thus experience daily threats to their identities. In developing their Collective Self-Esteem Scale (CSE), Luhtanen and Crocker (1992) provided an instrument that could help researchers address this gap.

The CSE was developed as a measure that could be used with any social, or collective identity; identities that are based on group memberships. Thus, the item content was not specific to the experiences of any one particular social group (e.g., ethnicity, gender). The 16-item CSE contains four four-item subscales: Membership, which assesses how involved they are with the groups; Private, which assess individuals' personal feelings about their group; Public, which corresponds to how individuals feel other people view their group; and Identity, which assess the degree to which the group membership is incorporated into their self-concept.

Although the CSE has been a widely used instrument for ethnic identity assessment, its development as a general scale for all collective identities is an inherent weakness for this purpose. Although the psychometric properties are generally sound for the scales used with far-ranging identities, one would be hard-pressed to argue that the dimensionality and meaning of these identities is equivalent. Sellers and colleagues attempted to address this limitation by developing the Multidimensional Model of Racial Identity (MMRI) and the corresponding Multidimensional Inventory of Black Identity (MIBI; Sellers, Rowley, Chavous, Shelton, & Smith, 1997; Sellers, Smith, Shelton, Rowley, & Chavous, 1998).

Influenced heavily by the CSE as well as other models of ethnic identity (e.g., Phinney's developmental model, Cross's nigrescence model), the MIBI assesses three dimensions of ethnic identity: centrality, regard, and ideology. The centrality dimension is akin to the Identity subscale of the CSE, assessing the degree to which individuals define themselves according to their ethnic background. It is also very similar to the Resolution subscale of the EIS or the Clarity subscale produced by some factor analyses of the MEIM-14 (Juang & Nguyen, 2010; Lee & Yoo, 2004).

The regard dimension contains two subscales, Private Regard and Public Regard. The distinction made by Sellers and colleagues between the types of regard flows directly from the CSE instrument. Although private regard is a construct that is similar to what is contained in other measures (i.e., affirmation in the EIS, commitment in the MEIM), the construct of public regard is unique to the CSE/MIBI measurement approach to ethnic identity.

The final dimension of the model, ideology, is perhaps the most unique contribution of the MIBI. Whereas the centrality and regard dimensions are not specific to the experiences of Black Americans, the ideology dimension is. Ideology comprises four subscales that correspond to four prevailing philosophies within Black American culture: Nationalist, Oppressed Minority, Assimilationist, and Humanist. Unfortunately, these subscales are rarely used.

The strength of the MIBI lies in its multidimensionality, particularly the distinction between private and public regard, and its inclusion of ethnicity-general (centrality, regard) and ethnicity-specific (ideology) dimensions. The primary weakness of the MIBI is that its psychometric properties have not been interrogated. As described by Vandiver, Worrell, and Delgado-Romero (2009) there has been very little psychometric analyses conducted since its original development and subsequent publication in 1997 (Sellers et al., 1997). The investigations that have been done yielded inconsistent findings, with the one consistency being that none have supported the factor structure proposed by Sellers et al. (1997).

A second limitation of the MIBI is that it is not firmly based within developmental theory. The volume of research based on the developmental model of ethnic identity, as measured by the various versions of the MEIM (e.g., Phinney, 1992), has clearly

demonstrated that ethnic identity is a developmental process that changes over time and context (French, Seidman, Allen, & Aber, 2006; Pahl & Way, 2006; Syed & Azmitia, 2009). None of the dimensions included in the MIBI have developmental properties, and thus no hypotheses about potential change over time can be derived. This limitation does not necessarily render the MIBI a weak instrument, but it does surface theoretical debates about the nature of ethnic identity that are far from resolved. The theoretical foundations of these various measures must be well-understood when selecting a measure to use.

Stage Theories of Ethnic Identity

The final class of ethnic identity models considered in this section of this chapter are stage theories, the most well known of which is Cross's (1971) nigrescence theory. To be clear, nigrescence theory is the earliest, and perhaps most influential, theory of ethnic identity. However, there is little to no empirical support for the model, which is why it is receiving relatively brief treatment here. Indeed, Nigrescence theory is probably the strongest and most clearly articulated theoretical model of ethnic identity, but also lays claim to having the weakest empirical support.

Nigrescence theory was developed as a means for describing the process through which individuals develop a socially conscious Black identity. The theory emerged in the wake of the civil rights movement, and thus has an inherently political and socially situated structure. The theory has been both revised and expanded since the original version (see Vandiver, Cross, Worrell, & Fhagen-Smith, 2002, for a chronicle of this progression). The current expanded nigrescence theory specifies four stages through which Black identity develops, each of which corresponds to an overarching theme that informs potential identities (Cross & Vandiver, 2001). The first stage, preencounter, indicates a low awareness of one's Black identity that may manifest as viewing being Black as not important, harboring negative attitudes toward other Blacks, or holding negative views of one's own Black identity. The next stage, encounter, suggests a period of awakening to one's Black identity, either through a singular event or series of events that prompts increased consciousness. Indeed, the encounter stage is less of a stage, per se, than it is a catalyst for moving the individual from the pre-encounter stage to the third stage, immersion–emersion. The immersion–emersion stage is the period through which individuals truly develop Black identities, the content of which is characterized by either an intensively overidentified. Black identity or a primarily anti-White identity. Ultimately individuals progress to the fourth and final stage, internalization, in which they develop a positive and salient Black identity paired with activism within the Black community (nationalist identity) or in coalition with other groups (biculturalist or multiculturalist identities).

The Cross Racial Identity Scale (CRIS) was developed to operationalize the expanded Nigrescence theory by assessing both the four stages and their corresponding identities (Vandiver et al., 2002). The model specifies nine stage–identity pairings, but the measure only assesses six of the nine, so it does not actually operationalize the theory. Moreover, without longitudinal analyses there is no way of knowing whether the subscales assess momentary structures of identity, stable styles of identity, or stages of development as the theory proposes. Indeed, a larger question looms about the applicability of stage theories at all for ethnic identity research, as recent research has documented the extreme variability in how and when individuals accelerate the process of ethnic identity development (Syed, 2010a).

Influenced by Nigrescence theory, Helms developed the White Racial Identity Attitude Scale (WRIAS; Helms & Carter, 1990) that was meant to operationalize a stage theory of White identity development. The theory postulated a six-stage developmental sequence: contact, disintegration, reintegration, pseudo-independence, immersion–emersion, and autonomy. Although provocative, repeated inquiries into the scales' psychometric properties have not provided evidence for the six factors (e.g., Behrens, 1997; Mercer & Cunningham, 2003; Swanson, Tokar, & Davis, 1994), and no longitudinal studies have been conducted to support the progression through the six stages. Helms's model of White identity development was the first and most detailed account available, which is why it has been included here, but the serious questions

about how well the WRIAS operationalizes the theory suggest that it should not be used (but see Helms, 2005, in defense of the measure).

In conclusion, there is little to no empirical support for stage models of ethnic identity development. The CRIS, although a psychometrically sound instrument, does not clearly or fully operationalize nigrescence theory. Thus, what the subscales mean is not clear. The weak link between the theories and measures, as well as the limitations of the measures themselves, suggest that these measures should generally be avoided. At the same time, these theories are the most theoretically rich and have the potential to generate the most theory-driven empirical research. The ethnic-specific focus of these theories and measures is a strength, but the nontransferability to other groups has limited their use and appeal for researchers who, rightly or wrongly, want to use the measures with other groups (unlike the MIBI, which has components that are easily modified for use with non-Blacks). The true fate of these theories, however, rests on the production of longitudinal research.

CHALLENGES TO ASSESSMENT OF ETHNIC IDENTITY

The many challenges to the field of ethnic identity will likely keep researchers busy for many years to come. Despite the abundance of research on ethnic identity over the past 20 to 25 years, comparatively little research has been devoted to issues of assessment. Three broad areas for future work are identified here: measurement equivalence, the meaning of ethnic identity, and theoretical debates.

Equivalence of ethnic identity measures across different ethnic groups has hardly been examined. With the exception of Roberts et al., (1999), most inquires into the factor structure of the MEIM have been conducted with rather small multiethnic samples (e.g., Gaines et al., 2010) or monoethnic samples that precluded analysis by ethnic group (e.g., Juang & Nguyen, 2010; Lee & Yoo, 2004). The latter class of studies afforded an opportunity to assess the structure along factors within groups (e.g., gender, immigrant generation status, socioeconomic status), but the small samples in these studies did not permit such tests.

Researchers who have found a three-factor solution for the MEIM-14 suggest that ethnic identity may be more differentiated as individuals develop; thus, a two-factor structure in early adolescence become a three-factor structure in young adulthood (Juang & Nguyen, 2010; Lee & Yoo, 2004). This suggestion is an intriguing proposal, but support for it could only be derived from testing for structural invariance over time in longitudinal studies. For example, Syed and Azmitia (2009) found evidence for invariance of the two-factor structure of the MEIM-12 over four years of college, but no such analyses were conducted in longitudinal studies by Pahl and Way (2006) or French et al. (2006).

Related to the issues of measurement equivalence of ethnic identity measures is the construct of meaning. Meaning is notably lacking from most measures of ethnic identity. Although it is captured by the Ideology subscales of the MIBI, these subscales are rarely used in research. Within ethnic minority populations, the meaning of ethnic identity may not be the same for immigrant and nonimmigrant youth, particularly in the United States. Ethnic identity, as conceptualized by Phinney's developmental model or Sellers's MMRI, is inherently a minority identity. The theory and measurement presumes that the respondent is located within a racially and ethnically stratified society. The subordinated position of ethnic minorities in the United States gives rise to the need for exploration, as ethnic minority youth often do not learn about their cultural background in schools or the larger society. In contrast, many countries around the world are relatively homogenous in terms of ethnicity, and in such contexts, ethnic identity may not hold much meaning. It is not until immigrants shift from a majority to minority context that ethnic identity becomes relevant. Still, at that point, exploring the meaning of one's ethnic background would not be a sensible practice for immigrants.

The meaning of ethnic identity for people who are in the majority, such as Whites in the United States, is also not well understood. Ethnic identity is believed to hold more meaning for ethnic minorities than ethnic majorities. Indeed, several studies have administered the MEIM to White youth, who have consistently scored lower on ethnic identity than do

youth from ethnic minority groups. To the knowledge of this chapter's author, however, as of 2010, there have not been any studies looking into how Whites interpret the items in the instrument. As described previously, Helms's model of White racial identity is one of the few available models of White identity, which is unfortunate, given its lack of established validity. White identity should be a major focus of research for years to come, particularly because the proportion of Whites in the United States is steadily decreasing, and in some regions they will be in the minority relative to ethnic minorities.

Perhaps the biggest challenge to ethnic identity assessment is the rising number of people who identify with multiple ethnic backgrounds. Phinney (1990) identified this as a major concern for ethnic identity research over 20 years ago, and unfortunately not much progress has been made. This concern is due, in part, to researchers' overemphasis on the labels that people from mixed-ethnic backgrounds choose rather than exploring what it means to have a mixed-ethnic identity. There is a recent movement to conceptualize mixed-ethnics as their own ethnic group, with such researchers arguing that they have more in common with each other than they do with their monoethnic peers (Syed, 2010a, 2010b; Syed & Azmitia, 2008). The number of mixed-ethnics is the United States is rapidly increasing (Lee & Bean, 2004), and thus this matter deserves immediate and thoughtful attention.

At the core of current challenges in assessing ethnic identities are theoretical debates that have yet to be resolved. The field of ethnic identity began as ethnic-specific investigations using instruments designed for the group in question (Phinney, 1990). The introduction of the MEIM-14 (Phinney, 1992) and the CSE (Luhtanen & Crocker, 1992) steered the ethnic identity literature in the direction of universal, panethnic models. The development of the MMRI and MIBI by Sellers and colleagues marked a return to the ethnic-specific approach. There is currently no consensus as to what the optimal model is or should be. Some have borrowed a page from cultural psychology, suggesting that the processes of ethnic identity may be universal, but the content of ethnic identity is ethnic specific (Syed & Azmitia, 2010).

The second theoretical debate that is inhibiting advances in assessment is the situational nature of ethnic identities on the one hand, and the developmental nature of ethnic identity on the others. The CSE, which operationalized a situational view on identities, was published in the same year (1992) as the MEIM-14, an operationalization of the developmental model. Each measure led to an explosion of subsequent research that is perhaps best characterized as two parallel explosions that rarely come into contact with one another. Only recently are researchers taking up the question of how ethnic identities may be both situational and follow a particular course of development. Initial examinations of this question by Yip and Fuligni (2002) found that Chinese American adolescents who were actively engaging in ethnicity-related behaviors reported greater ethnic identity salience at that moment. Moreover, ethnic identity salience was only associated with well-being for those adolescence who reported stronger ethnic identities. This study made a significant contribution, as it linked measurement at the situational level with measurement of more stable elements of identity. It was limited, however, in that it was conducted over a very brief time period (2 weeks). Thus, how situational behaviors and salience are related to the development course of identity remains unknown. Such investigations would make a welcome contribution to the assessment of ethnic identity.

THEORIES AND MEASURES OF ACCULTURATION

Acculturation broadly refers to the process of adaptation in response to sustained[2] intercultural contact (Berry, Phinney, Sam, & Vedder, 2006). Thus, in contrast to ethnic identity, which focuses on individuals' self-realization of the importance of their own cultural heritage, acculturation is concerned

[2]The word *sustained* is a key part of this definition. Acculturation theories are not believed to be appropriate for temporary or transitory intercultural contact, as in the case of travel, sojourners, international students, or so-called third-culture kids. For these groups, theories of cultural adaptation may be more relevant (see Ward & Kennedy, 1999).

with how individuals negotiate identities when two or more cultures come in contact. Acculturation research is primarily focused on immigrant populations, who must reconcile the beliefs, values, and practices of their culture of origin with those of their host culture. Because of the ever-increasing amount of immigration in countries throughout the world, acculturation research has a much greater international focus than ethnic identity. Within the United States, however, acculturation research has been confined mostly to immigrants from Asia and Latin America.

There are two general theoretical views on how the process of acculturation unfolds over time (see Juang & Nguyen, 2011, and Nguyen, Messé, & Stollak, 1999, for more discussion). One view, referred to as the unidimensional or bipolar approach, conceptualizes the two cultures as lying on a single continuum, mutually dependent and forever in tension. For a Chinese immigrant living in the United States, increased identification with U.S. culture necessitates a decreased identification with Chinese culture. Under this model, adaptation is conceptualized as assimilation on the one hand (adopting U.S. culture) or separation on the other (maintaining Chinese culture). This view of acculturation is consistent with the metaphor of the melting pot in the United States, wherein immigrants gradually shed their culture of origins and blend in with other Americans.

The bidimensional view of acculturation stands in stark contrast to the unidimensional approach. Rather than specifying a single mutually dependent cultural continuum, proponents of the bidimensional approach argue that involvement with two cultures is best operationalized as orthogonal (Berry et al., 2006). That is, acculturation is a process of simultaneous identification with one's culture of origin *and* one's host culture, and these two dimensions are theoretically independent. Under this view, it is possible for an individual to have a strong identification with both cultures. The independence of the two dimensions has been demonstrated in multiple studies, providing strong empirical support for the bidimensional model (e.g., Cheung-Blunden & Juang, 2008; Miller, 2010; Ryder, Allen, & Paulhus, 2000). This view of acculturation is consistent with the salad bowl metaphor in the United States and the cultural mosaic in Canada, in which immigrants are able to preserve the integrity of their cultural "ingredient" while still being a part of the larger salad. Thus, the bidimensional model of acculturation is more aligned with the ideals of a just, multicultural society, and has therefore been adopted as the currently preferred approach in the field. In the words of Juang and Nguyen (2011), one of the major advantages of bidimensional assessment is that it "allows researchers to test whether the two dimensions relate to one another, relate differentially to adjustment, and/or interact with one another in relating to adjustment" (p. 73). In other words, bidimensional assessment affords much greater analytic options than with unidimensional assessment (Rudmin, 2009).

Adopting a definition of acculturation as broad as adaptation in response to sustained cultural contact opens the door to a wide variety of domains to investigate. Indeed, under the banner of acculturation, researchers have investigated endorsements of attitudes, values, beliefs, behaviors, and language use (Zane & Mak, 2003). Accordingly, many measures of acculturation are currently in use. In their review of 21 acculturation measures, Zane and Mak (2003) found very little consistency across instruments (see also Wallace et al., 2010, for a review of measures used with Latinos in the United States). The conceptual overlap among the measures was minimal, and the range of domains assessed was neither consistent across measures nor extensive within a measure. Heritage language use was by far the most dominant indicator of acculturation. Surprisingly, they also found that 14 of the 21 measures (67%) assessed a unidimensional rather than bidimensional model of acculturation. Indeed, despite the perceived superiority of the bidimensional model, there are surprisingly few instruments that adequately assess acculturation bidimensionally.

In comparison with ethnic identity research, the connection between assessment instruments and the theory they are meant to operationalize within acculturation research is rather weak (see also Rudmin, 2009). From Zane and Mak's (2003) review, one can conclude that a measure that samples across

a variety of domains *and* operationalizes a bidimensional theory of acculturation is quite rare, and thus advances in assessment have generally not stayed in line with advances with theory. There are a few measures that meet these standards, which are described in the following text. The upshot of this limitation is that those interested in theory-driven assessment need not wade through the myriad choices of acculturation measures, given that so few adequately assess the theory.

The Acculturation Rating Scale for Mexican Americans–II (ARSMA-II) is one of the most widely used measures of acculturation of Mexican-Americans (Cuéllar, Arnold, & Maldonado, 1995). The 30-item ARSMA-II comprises two subscales, 17 items that pertain to orientation toward Mexican culture and 13 items that pertain to orientation toward American culture. The items sample across multiple domains, including language, identity, and behaviors. By having two distinct scales for Mexican and American involvement, the ARSMA-II appropriately operationalizes a bidimensional model of acculturation. It is important to note that the original scale, the ARSMA, was composed of a single scale that operationalized a linear model of acculturation, and therefore is not recommended. The ARSMA-II has been used with other Latino-heritage ethnic groups and has been translated into Spanish.

Like the ARSMA-II, the Acculturation Scale for Vietnamese Adolescents (ASVA) assesses a bidimensional acculturation structure across a variety of domains (Nguyen & von Eye, 2002). The 50-item instrument comprises two 25-item subscales, Involvement in the Vietnamese Culture and Involvement with the American culture. The two subscales contain questions pertaining to attitudes, behaviors, and values across four domains: everyday lifestyles, group interactions, family orientation, and global involvement. The four domains make up subscales of their own within each cultural orientation (Vietnamese or American), resulting in a total of eight possible subscales nestled within two higher order factors. This factor structure was supported by a series of confirmatory factor analyses, and the authors also demonstrated through model comparisons that a bidimensional model was superior to a unidimensional model. This measure has also been modified to be used with other Asian and Asian American samples (Cheung-Blunden & Juang, 2008).

CHALLENGES TO ASSESSMENT OF ACCULTURATION

Many of the challenges described in ethnic identity assessment are equally applicable to assessing acculturation. For example, whether the process of acculturation is best conceptualized as universal across ethnic groups or ethnic-group specific remains an important question. The two measures described previously, the ARSMA-II and the AVSA were both developed as ethnic-specific instruments, but then modified and adopted for other groups (much like the MIBI in ethnic identity research).

Because of the breadth of content covered by the term *acculturation*, a major question that looms in the field is whether instruments that assess acculturation as a universal, cross-domain process are appropriate. Reflecting a movement toward domain-specific assessment, recent research situated within bidimensional acculturation theory has examined family distancing, family conflict, loss of face, parental control, and bicultural identity as specific components of acculturation (Benet-Martínez & Haritatos, 2005; Hwang & Wood, 2009; Juang, Syed, & Takagi, 2007; Lee, Choe, Ngo, & Kim, 2000; Zane & Mak, 2003).

In assuming a critical acculturation stance, Rudmin (2009) has painted a gloomy picture of the past and present acculturation research and calls for a complete overhaul of the field, going so far as to say, "nearly one century of acculturation research has resulted in little reliable or useful information" (p. 108). This assertion is based on the measurement limitations described earlier, the overemphasis on the effects of acculturation on health, and questions about whether there is as optimal way to acculturate. Some of these arguments stand on flimsy ground (Berry, 2009), and no assessment tools have been advanced to supplant the existing ones. Thus, critical acculturation is not currently of major concern to the field. Nevertheless, in a field dominated by a dimensionality debate, the challenges brought by critical acculturation may help move the field

forward and bring greater coherence to acculturation research.

CONCLUSION

Although research on ethnic identity and acculturation has been conducted, in some form, for nearly 100 years (Rudmin, 2009), the fields are rather nascent in terms of rigorous, cumulative research. The vast majority of measurement tools are idiosyncratic and atheroretical. The motivation for including in this chapter only those instruments that have strong theoretical foundations is to encourage future research to build from these existing strengths to increase the cumulativeness of the fields. Contrary to Rudmin's assertions, psychology has actually learned a great deal of reliable and useful information, particularly over the past 20 to 25 years, much of which is due to the valuable research covered in this chapter. However, there is still much work to do, and if we are truly invested in how cultural identities matter in people's lives, we would do well to spend greater time and attention to how we assess those identities.

References

Ashmore, R. D., Deaux, K., & McLaughlin-Volpe, T. (2004). An organizing framework for collective identity: Articulation and significance of multidimensionality. *Psychological Bulletin*, *130,* 80–114. doi:10.1037/0033-2909.130.1.80

Behrens, J. T. (1997). Does the White Racial Identity Attitude Scale measure racial identity? *Journal of Counseling Psychology*, *44,* 3–12. doi:10.1037/0022-0167.44.1.3

Benet-Martínez, V., & Haritatos, J. (2005). Bicultural Identity Integration (BII): Components and psychosocial antecedents. *Journal of Personality, 73*, 1015–1050. doi:10.1111/j.1467-6494.2005.00337.x

Berry, J. W. (2009). A critique of critical acculturation. *International Journal of Intercultural Relations*, *33,* 361–371. doi:10.1016/j.ijintrel.2009.06.003

Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006). Immigrant youth: Acculturation, identity, and adaptation. *Applied Psychology, 55,* 303–332. doi:10.1111/j.1464-0597.2006.00256.x

Cheung-Blunden, V. L., & Juang, L. P. (2008). Expanding acculturation theory: Are acculturation models and the adaptiveness of acculturation strategies generalizable in a colonial context? *International Journal of Behavioral Development*, *32,* 21–33. doi:10.1177/0165025407084048

Cross, W. E. (1971). The Negro to Black conversation experience: Toward a psychology of Black liberation. *Black World*, *20,* 13–27.

Cross, W. E., Jr., & Vandiver, B. J. (2001). Nigrescence theory and measurement: Introducing the Cross Racial Identity Scale (CRIS). In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (2nd ed., pp. 371–393). Thousand Oaks, CA: Sage.

Cuéllar, I., Arnold, B., & Maldonado, R. (1995). Acculturation rating scale for Mexican Americans—II: A revision of the original ARSMA scale. *Hispanic Journal of Behavioral Sciences*, *17,* 275–304. doi:10.1177/07399863950173001

Erikson, E. H. (1968). *Identity: Youth and crisis.* New York, NY: Norton.

French, S. E., Seidman, E., Allen, L., & Aber, J. L. (2006). The development of ethnic identity during adolescence. *Developmental Psychology, 42*, 1–10. doi:10.1037/0012-1649.42.1.1

Gaines, S. O., Jr., Bunce, D., Robertson, T., Wright, B., Goossens, Y., Heer, D., . . . Minhas, S. (2010). Evaluating the psychometric properties of the Multigroup Ethnic Identity Measure (MEIM) within the United Kingdom. *Identity: An International Journal of Theory and Research*, *10,* 1–19. doi:10.1080/15283481003676176

Helms, J. E. (2005). Challenging some misuses of reliability as reflected in evaluations of the White Racial Identity Attitude Scale (WRIAS). In R. T. Carter (Ed.), *Handbook of racial–cultural psychology and counseling: Vol. 1. Theory and research* (pp. 360–390). Hoboken, NJ: Wiley.

Helms, J. E., & Carter, R. T. (1990). Development of the White Racial Identity Inventory. In J. E. Helms (Ed.), *Black and White racial identity attitudes: Theory, research and practice* (pp. 67–80). Westport, CT: Greenwood Press.

Hwang, W. C., & Wood, J. J. (2009). Acculturative family distancing: Links with self-reported symptomatology among Asian Americans and Latinos. *Child Psychiatry and Human Development*, *40,* 123–138. doi:10.1007/s10578-008-0115-8

Jackson, Y. (Ed.). (2006). *Encyclopedia of multicultural psychology*. Thousand Oaks, CA: Sage.

Juang, L. P., & Nguyen, H. H. (2010). Ethnic identity among Chinese-American youth: The role of family obligation and community factors on ethnic engagement, clarity, and pride. *Identity: An International Journal of Theory and Research*, *10,* 20–38. doi:10.1080/15283481003676218

Juang, L. P., & Nguyen, H. H. (2011). Acculturation and adjustment in Asian American children and families.

In F. Leong, L. P. Juang, D. Qin, & H. Fitzgerald (Eds.), *Asian American child psychology and mental health* (pp. 71–95). Westport, CT: Praeger.

Juang, L. P., Syed, M., & Takagi, M. (2007). Intergenerational discrepancies of parental control among Chinese American families: Links to family conflict and adolescent depressive symptoms. *Journal of Adolescence*, *30,* 965–975. doi:10.1016/j.adolescence.2007.01.004

Landis, D., Bennett, J. M., & Bennett, M. J. (Eds.). (2004). *Handbook of intercultural training* (3rd ed.). Thousand Oaks, CA: Sage.

Lee, J., & Bean, F. D. (2004). America's changing color lines: Immigration, Race/Ethnicity, and multiracial identification. *Annual Review of Sociology, 30*, 221–242. doi:10.1146/annurev.soc.30.012703.110519

Lee, R., Choe, J., Ngo, G., & Kim, V. (2000). Construction of the Asian American family conflicts scale. *Journal of Counseling Psychology, 47*, 211–222. doi:10.1037/0022-0167.47.2.211

Lee, R. M., & Yoo, H. C. (2004). Structure and measurement of ethnic identity for Asian American college students. *Journal of Counseling Psychology*, *51,* 263–269. doi:10.1037/0022-0167.51.2.263

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin*, *18,* 302–318. doi:10.1177/0146167292183006

Marcia, J. E. (1966). Development and validation of ego identity status. *Journal of Personality and Social Psychology, 3*, 551–558. doi:10.1037/h0023281

Marcia, J. E. (1980). Identity in adolescence. In J. Adelson (Ed.), *Handbook of adolescent psychology* (pp. 159–197). New York, NY: Wiley.

Markus, H. R. (2008). Pride, prejudice, and ambivalence: Toward a unified theory of race and ethnicity. *American Psychologist*, *63,* 651–670. doi:10.1037/0003-066X.63.8.651

Mercer, S. H., & Cunningham, M. (2003). Racial identity in White American college students: Issues of conceptualization and measurement. *Journal of College Student Development*, *44,* 217–230. doi:10.1353/csd.2003.0021

Miller, M. J. (2010). Testing a bilinear domain-specific measure of acculturation and enculturation across generational status. *Journal of Counseling Psychology, 57*, 179–186. doi:10.1037/a0019089

Nguyen, H. H., Messé, L. A., & Stollak, G. E. (1999). Toward a more complex understanding of acculturation and adjustment. *Journal of Cross-Cultural Psychology*, *30,* 5–31. doi:10.1177/0022022199030001001

Nguyen, H. H., & von Eye, A. (2002). The acculturation scale for Vietnamese adolescents (ASVA): A bidimensional perspective. *International Journal of Behavioral Development*, *26,* 202–213. doi:10.1080/01650250042000672

Pahl, K., & Way, N. (2006). Longitudinal trajectories of ethnic identity among urban Black and Latino adolescents. *Child Development*, *77,* 1403–1415. doi:10.1111/j.1467-8624.2006.00943.x

Phinney, J. S. (1989). Stages of ethnic minority development in minority group adolescents. *Journal of Early Adolescence*, *9,* 34–49. doi:10.1177/0272431689091004

Phinney, J. S. (1990). Ethnic identity in adolescents and adults: A review of research. *Psychological Bulletin*, *108*, 499–514. doi:10.1037/0033-2909.108.3.499

Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research*, *7,* 156–176. doi:10.1177/074355489272003

Phinney, J. S. (1993). A three-stage model of ethnic identity in adolescence. In M. E. Bernal & G. P. Knight (Eds.), *Ethnic identity: Formation and transmission among Hispanics and other minorities* (pp. 61–79). Hillsdale, NJ: Erlbaum.

Phinney, J. S., & Ong, A. D. (2007). Conceptualization and measurement of ethnic identity: Current status and future directions. *Journal of Counseling Psychology*, *54,* 271–281. doi:10.1037/0022-0167.54.3.271

Quintana, S. M. (2007). Racial and ethnic identity: Developmental perspectives and research. *Journal of Counseling Psychology*, *54,* 259–270. doi:10.1037/0022-0167.54.3.259

Roberts, R. E., Phinney, J. S., Masse, L. C., Chen, Y. R., Roberts, C. R., & Romero, A. (1999). The structure of ethnic identity of young adolescents from diverse ethnocultural groups. *Journal of Early Adolescence*, *19,* 301–322. doi:10.1177/0272431699019003001

Rudmin, F. (2009). Constructs, measurements and models of acculturation and acculturative stress. *International Journal of Intercultural Relations*, *33,* 106–123. doi:10.1016/j.ijintrel.2008.12.001

Ryder, A. G., Alden, L. E., & Paulhus, D. L. (2000). Is acculturation unidimensional or bidimensional? A head-to-head comparison in the prediction of personality, self-identity, and adjustment. *Journal of Personality and Social Psychology, 79*, 49–65. doi:10.1037/0022-3514.79.1.49

Seaton, E. K., Scottham, K. M., & Sellers, R. M. (2006). The status model of racial identity development in African American adolescents: Evidence of structure, trajectories, and well-being. *Child Development*, *77,* 1416–1426. doi:10.1111/j.1467-8624.2006.00944.x

Sellers, R. M., Rowley, S. A. J., Chavous, T. M., Shelton, J. N., & Smith, M. A. (1997). Multidimensional inventory of Black identity: A preliminary investigation of reliability and construct validity. *Journal*

of Personality and Social Psychology, *73,* 805–815. doi:10.1037/0022-3514.73.4.805

Sellers, R. M., Smith, M. A., Shelton, J. N., Rowley, S. A. J., & Chavous, T. M. (1998). Multidimensional model of racial identity: A reconceptualization of African American racial identity. *Personality and Social Psychology Review*, *2,* 18–39. doi:10.1207/s15327957pspr0201_2

Swanson, J. L., Tokar, D. M., & Davis, L. E. (1994). Content and construct validity of the white racial identity attitude scale. *Journal of Vocational Behavior*, *44,* 198–217. doi:10.1006/jvbe.1994.1014

Syed, M. (2010a). Developing an integrated self: Academic and ethnic identities among ethnically diverse college students. *Developmental Psychology*, *46,* 1590–1604. doi:10.1037/a0020738

Syed, M. (2010b). Memorable everyday events in college: Narratives of the intersection of ethnicity and academia. *Journal of Diversity in Higher Education*, *3,* 56–69. doi:10.1037/a0018503

Syed, M., & Azmitia, M. (2008). A narrative approach to ethnic identity in emerging adulthood: Bringing life to the identity status model. *Developmental Psychology*, *44,* 1012–1027. doi:10.1037/0012-1649.44.4.1012

Syed, M., & Azmitia, M. (2009). Longitudinal trajectories of ethnic identity during the college years. *Journal of Research on Adolescence*, *19,* 601–624. doi:10.1111/j.1532-7795.2009.00609.x

Syed, M., & Azmitia, M. (2010). Narrative and ethnic identity exploration: A longitudinal account of emerging adults' ethnicity-related experiences. *Developmental Psychology*, *46,* 208–219. doi:10.1037/a0017825

Syed, M., Azmitia, M., & Phinney, J. S. (2007). Stability and change in ethnic identity among Latino emerging adults in two contexts. *Identity: An International Journal of Theory and Research*, *7,* 155–178. doi:10.1080/15283480701326117

Tajfel, H. (1981). *Human groups and social categories.* Cambridge, England: Cambridge University Press.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1,* 149–178. doi:10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Umaña-Taylor, A. J. (2004). Ethnic identity and self-esteem: Examining the role of social context. *Journal of Adolescence, 27*, 139–146. doi:10.1016/j.adolescence.2003.11.006

Umaña-Taylor, A. J., Yazedjian, A., & Bámaca-Gómez, M. (2004). Developing the Ethnic Identity Scale using Eriksonian and social identity perspectives. *Identity, 4,* 9–38.

Vandiver, B. J., Cross, W. E., Jr., Worrell, F. C., & Fhagen-Smith, P. E. (2002). Validating the Cross Racial Identity Scale. *Journal of Counseling Psychology*, *49,* 71–85. doi:10.1037/0022-0167.49.1.71

Vandiver, B. J., Worrell, F. C., & Delgado-Romero, E. A. (2009). A psychometric examination of multidimensional inventory of Black identity (MIBI) scores. *Assessment, 16*, 337–351. doi:10.1177/1073191109341958

Wallace, P. M., Pomery, E. A., Latimer, A. E., Martinez, J. L., & Salovey, P. (2010). A review of acculturation measures and their utility in studies promoting Latino health. *Hispanic Journal of Behavioral Sciences*, *32,* 37–54. doi:10.1177/0739986309352341

Ward, C., & Kennedy, A. (1999). The measurement of sociocultural adaptation. *International Journal of Intercultural Relations*, *23,* 659–677. doi:10.1016/S0147-1767(99)00014-0

Yip, T., & Fuligni, A. J. (2002). Daily variation in ethnic identity, ethnic behaviors, and psychological well-being among American adolescents of Chinese descent. *Child Development*, *73,* 1557–1572. doi:10.1111/1467-8624.00490

Yip, T., Seaton, E. K., & Sellers, R. M. (2006). African American racial identity across the lifespan: Identity status, identity content, and depressive symptoms. *Child Development*, *77,* 1504–1517. doi:10.1111/j.1467-8624.2006.00950.x

Zane, N., & Mak, W. (2003). Major approaches to the measurement of acculturation among ethnic minority populations: A content analysis and an alternative empirical strategy. In K. Chun, P. Organista, & G. Marin (Eds.), *Acculturation: Advances in theory, measurement, and applied research* (pp. 39–60). Washington, DC: American Psychological Association. doi:10.1037/10472-005

CHAPTER 24

ASSESSMENT OF PERSONALITY IN COUNSELING SETTINGS

Margit I. Berman and Sueyoung L. Song

Personality assessment is a major activity of counseling psychologists. Although the American Psychological Association (APA) description of counseling psychology, which defines the specialty, does not mention personality assessment per se, it does imply that such assessment is part of the work of counseling psychologists. Specifically, it describes "assessment and diagnosis of psychopathology" as a basic element of the specialty, "psychological measurement and principles of psychological/diagnostic and environmental assessment" as areas of scientific knowledge germane to the specialty, and lists "psychodiagnostic assessment techniques" as among the procedures used by counselors (APA, n.d.). Counseling psychologists in various practice settings use a variety of both objective and projective personality tests (Fee, Elkins, & Boyd, 1982; Watkins & Campbell, 1989), such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951), the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Hathaway & McKinley, 1991), the Rorschach (Exner, 1969), and the Thematic Apperception Test (TAT; Murray, 1971). Courses in psychological assessment are required by most counseling psychology training programs, and many training programs require courses specifically in objective and/or projective personality testing. In addition, the vast majority of counseling psychology program directors report that training in personality assessment is of substantial or great importance for counseling students (Watkins, Campbell, & Manus, 1990).

Most of the specific personality tests and assessment tools most commonly used by counselors in practice, such as the MMPI/MMPI–2, the TAT, or sentence completion blanks, would be equally, if not more, familiar to clinical psychologists, raising the question of "How special is the [counseling] specialty" (Fitzgerald & Osipow, 1986)? Indeed, distinguishing a uniquely "counseling" approach to personality assessment can be difficult, given the diversity of counseling psychologists' research, training and practice roles. However, the unique strengths and values of the field are visible in many aspects of personality assessment as researched, taught, and used by counseling psychologists, and there is also evidence that counseling psychologists are making distinct contributions to the field of personality assessment as a whole.

In this chapter, we begin by considering "How special is the specialty?" and delineate both what is unique about counseling psychology's approach to personality assessment as well as what aspects of personality assessment counseling psychology shares with other psychologists (e.g., clinical psychologists). Next, we describe how counselors use personality assessment in practice; briefly review and describe how counselors assess personality, with attention to the most commonly used assessment methods; and discuss multicultural issues in personality assessment in counseling. Finally, we conclude with recommendations for research and counseling practice of personality assessment.

DOI: 10.1037/14048-024
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

PERSONALITY ASSESSMENT: WHAT MAKES COUNSELING PSYCHOLOGY UNIQUE?

Compared with clinical psychology, personality assessment in counseling psychology emphasizes normal (nonpathological) aspects of personality as assessed by objective (rather than projective) measures (Watkins, 1983; Watkins, Campbell, Hollifield, & Duckworth, 1989). Despite this general truism, however, there is broad variability in counselors' professional roles and activities, and it is likely that counselors use nearly all of the same personality assessment tools used by clinical psychologists. Nevertheless, even when using personality assessments strongly identified with clinical psychology, counselors may take a unique approach to the assessment process.

Duckworth (1990), for example, described a "counseling approach" to the use of psychological tests consisting of five distinct elements. First, this approach suggests that using personality tests may aid in short-term therapy, because testing may permit both more rapid problem conceptualization and treatment planning for the counselor and also, for clients, more rapid acquisition of insight into their own personalities. Second, counselors may use tests repeatedly across treatment to emphasize the developmental nature of client problems and the possibility of client change. Third, testing in counseling may help clients obtain a more accurate, objective picture of themselves, including their strengths as well as their weaknesses, that can be used by counselors as an aid to problem solving or in challenging a maladaptive self-concept. Fourth, counselors may optimize clients' ability to use test information to make decisions by involving clients in the testing process, such as by encouraging them to help choose what tests should be used, providing feedback directly to them rather than primarily to other professionals, and adjusting interpretation feedback to meet their decision-making needs. Fifth, counselors assume that clients will use test information adaptively, and therefore adopt a collaborative, psychoeducational style in providing assessment information to clients.

Although this "counseling approach" may initially appear simply to reiterate more general guidelines for use of assessment tools (e.g., the APA ethics code specifies that test takers receive feedback about their results unless the nature of the assessment relationship precludes it), in some ways this approach represents a significant departure from the ways in which clinical psychologists use personality tests. For example, Duckworth's (1990) emphasis on the development and change of personality during counseling, and the suggestion to use personality measures at various points in therapy to demonstrate change to the client, stands in contrast to trait theories of personality, which suggest that these tests measure traits which are stable and unlikely to change (Costa & McCrae, 1992b).[1]

In addition to the unique approach counselors may bring to the personality assessment process, some of the settings in which counselors work and the uses to which their assessments may be put may be relatively unique. Counselors may be more likely than other applied psychologists to work in college counseling centers, or to be involved in employee selection, organizational consulting, and life coaching. In these settings, using tests to optimize educational and occupational choices may be a major and more uniquely "counseling" role, as may be helping individuals, teams or organizations understand and use personality assessment information to enhance their performance.

The specific measures counselors use may also be a reflection of a uniquely counseling psychology approach to personality assessment. Some personality assessments, such as the California Personality Inventory (CPI; Gough, 2000) or the Myers–Briggs Type Indicator (MBTI; Myers & McCauley, 1985), although by no means used solely by counseling psychologists, are strongly identified with the specialty and represent the values of attention to human strengths, normal personality, and vocational issues that are typical of counseling psychology.

[1]There is surprisingly little recent research addressing the question of whether therapy or counseling actually can significantly alter personality assessment results, although early research in client-centered therapy suggested that this finding was possible (e.g., Gallagher, 1953). A few more recent longitudinal outcome studies have also demonstrated personality test score changes over the course of psychodynamic therapy (Itzhar-Nabarro, Silberschatz, & Curtis, 2009; Monsen, Odland, Faugli, Daae, & Eilertsen, 1995).

Counseling psychology's approach to personality assessment may also be distinguished by measures and techniques that are relatively rarely used, or de-emphasized in training and counseling practice, such as projective or neuropsychological assessment. Most counseling psychology training programs neither require nor offer electives in projective personality assessment (Watkins, Campbell, & Manus, 1990; see also Chapter 10 in this volume for additional information on projective or performance-based measures). Neuropsychology is another area which counseling psychology has traditionally de-emphasized. A minority of counseling psychology training programs offer any course in neuropsychology, and counseling psychology training directors express some reservations about how consistent neuropsychology is with counseling psychology as a discipline (Ryan, Lopez, & Lichtenberg, 1999).

Despite these unique qualities of personality assessment within counseling psychology in terms of approach, measures used or de-emphasized, and research and training, the "specialness of the specialty" with regard to personality assessment should not be overstated. Counseling psychologists routinely use a broad array of personality assessment techniques in research, practice, and training, including neuropsychological and projective measures. Counseling psychology is the second most common degree specialization among American Board of Professional Psychology/American Board of Clinical Neuropsychology diplomates (Ryan & Lopez, 1996), and a special issue of *The Counseling Psychologist* was devoted to counseling neuropsychology (Larson & Agresti, 1992). Similarly, many counseling psychologists have called for greater involvement of the specialty in projective personality assessment (Clark, 1995; Watkins, Campbell, Hollifield, et al., 1989). Counseling psychologists have also made substantial scholarly contributions to research and training of personality assessment beyond the development and validation of measures closely identified with the specialty. In a survey of the individuals who provided MMPI training in counseling psychology graduate programs, nearly half (46%) reported using the MMPI in their own research, and a substantial minority (31%) had published research on the MMPI (Watkins, Campbell, McGregor, & Godin, 1989).

In practice, counseling psychologists' use of assessment tools strongly resembles that of their clinical psychologist colleagues. Table 24.1 displays the "top 10" assessment procedures (not including

TABLE 24.1

Clinical and Counseling Psychologists' Top 10 Most Frequently Used Assessment Procedures

Counselors (Watkins & Campbell, 1989)	Clinicians (Watkins, Campbell, Nieberding, & Hallmark, 1995)
Strong-Campbell Interest Inventory	Wechsler Adult Intelligence Scale—Revised
Minnesota Multiphasic Personality Inventory	Minnesota Multiphasic Personality Inventory—2
Wechsler Adult Intelligence Scale—Revised	Sentence Completion Methods
Sentence Completion Blanks	Thematic Apperception Test
Bender–Gestalt	Rorschach
Thematic Apperception Test	Bender–Gestalt
16 Personality Factor Questionnaire	Projective Drawings
Wechsler Intelligence Scale for Children—Revised	Beck Depression Inventory
Draw-A-Person	Wechsler Intelligence Scale for Children—III
House-Tree-Person	Wide Range Achievement Test—Revised

Note. Data presented in this table are taken from surveys of counseling (Watkins & Campbell, 1989) and clinical psychologists' (Watkins, Campbell, Nieberding, & Hallmark, 1995) assessment practices. Although the methodologies used in both surveys were similar, there were some differences. Both groups were asked to rate the degree to which they used a preselected list of assessment procedures; these lists are not provided in either article but appear not to have been identical (e.g., "clinical interview" appears not to have been an option for the counselors and thus is omitted from this comparative table despite being the most commonly used measure for clinicians). Thus, the absence of a measure on either list is not definitive evidence that the group does not commonly use it, as it may not have been an option on the survey.

the clinical interview) reported by both counselors and clinical psychologists in practice; the lists are strikingly similar, both in terms of the measures themselves and the frequency with which psychologists report using them. Only a few measures are unique to one list or the other: the Rorschach, the Beck Depression Inventory (BDI; Beck & Steer, 1987), and the Wide Range Achievement Test (WRAT; Jastek & Wilkinson, 1984) on the clinicians' list, and the 16 Personality Factor Questionnaire (16PF; Cattell, Eber, & Tatsuoka, 1970) and Strong Campbell Interest Inventory (now called the Strong Interest Inventory; SII; Harmon, Hansen, Borgen, & Hammer, 1994) for the counselors. The presence of the Strong may represent the most important difference between counselors and clinicians; this measure was the most commonly used measure for counselors, but not in the top 10 for clinicians at all, although it is uncertain if the Strong was among the choices from which clinicians selected.

What specific assessment tools counselors should be trained to use, and whether or not there is a uniquely "counseling" approach to personality assessment, are questions which remain controversial, particularly in a training and practice environment where counseling psychologists' professional roles are increasingly diverse. Of even more concern is the general decline both within counseling psychology and in psychology as a whole in the use of and evaluation of the utility of personality assessment. The counseling psychology journals publish relatively little research on personality assessment, and neither the current nor the most recent previous edition of the *Handbook of Counseling Psychology* (Brown & Lent, 2000, 2008) features a chapter on personality assessment. As the reader may have observed in the brief review just completed, the literature on how counselors use assessment tools is also dated and limited. This lack of scholarly interest in personality assessment in counseling may not be a failing of counseling psychology specifically so much as a simple reflection of the changing landscape for personality assessment in psychology as a whole. The value of personality assessment as a psychological activity and whether assessment is in decline or thriving have been intensely debated by psychologists in recent years, as the use of assessment in applied psychology has been sharply challenged, while thriving in other areas of psychology, such as industrial–organizational psychology (Eisman et al., 2000; Meyer et al., 2001).

Despite these controversies and issues both within counseling psychology and for the practice of personality assessment as a whole, personality assessment remains an important activity for counseling psychologists in diverse professional roles and settings. Counseling psychologists use assessment tools for many of the same purposes as other psychologists do, and some purposes more unique to the specialty. The diverse ways counseling psychologists use personality assessment tools are considered in detail next.

HOW COUNSELORS USE PERSONALITY ASSESSMENT

As scientist-practitioners, counseling psychologists are involved in multiple roles, including helping clients and organizations, teaching, research, advocacy, and prevention. Counseling psychologists also work with diverse client populations in discharging these roles. Counseling psychologists see individuals with severe and persistent mental or physical illness requesting personality assessment to aid in basic vocational rehabilitation; they also see business executives requesting personality assessment to maximize top-level productivity in their organizations. The use of personality assessments by counseling psychologists thus may vary depending on settings and clients. Counseling psychologists use personality assessment in many of the same settings and for the same purposes as other psychologists, but there are also some settings and purposes that are more closely identified with counseling. Across settings, client types, and professional activities, counseling psychologists use personality assessment for a variety of key purposes, including (a) to help clients make better life choices across their development, (b) to improve decision making and performance in organizations, (c) to improve the outcome of counseling and psychotherapy, and (d) in research. Each of these purposes is briefly reviewed here.

MAKING BETTER LIFE CHOICES

Counseling psychologists use personality assessments across the lifespan of their clients to enhance decision making at critical developmental points, often in conjunction with other interventions and other assessment types (e.g., assessment of values or decision-making style). College counseling centers are major users of personality and other assessments, but they appear to use assessments in a fashion that distinguishes them from other settings (e.g., community mental health clinics or psychiatric hospitals; Lubin, Larsen, Matarazzo, & Seever, 1985). Although counseling centers, as with counseling psychologists as a whole, use the SII more commonly than any other psychological test, personality assessments, especially the MMPI–2, are also popular. Personality assessments may be used in college counseling centers to improve college major choice, a vital decision for college students' adjustment and future vocational plans, and one for which person–environment fit may be strongly predicted by personality measures (Porter & Umbach, 2006). In addition, counseling psychologists use personality assessments in career counseling, to assist with vocational choice. Individualized interpretation and feedback, often of test information, has been identified as one of five critical components that improve the outcome of career counseling interventions (Brown & Ryan Krane, 2000).

Counseling psychologists also use personality assessment to enhance relationship functioning in couples and families. The MBTI has been frequently used with couples to help them develop complementary styles of relating to one another and to understand their differences and similarities (Williams & Tappan, 1995). Similarly, the 16PF can be purchased as a Couples Counseling Form (Snyder, 1997), with questionnaires for both partners and supplementary items assessing the satisfaction of the relationship.

Counseling psychologists have also used personality assessment in innovative ways to enhance life choices made by clients. For example, one study examined relations among personality, self-efficacy, and stages of change among college students engaging in physical exercise, offering suggestions to counselors about how to use knowledge about personality traits to enhance the effectiveness of interventions designed to encourage students to exercise (Buckworth, Granello, & Belmore, 2002).

IMPROVING PERFORMANCE IN ORGANIZATIONS

Personality assessment has long been a major work activity for counseling psychologists who consult with organizations, provide executive coaching, or are involved in personnel selection or decision making. Although more strongly associated with industrial–organizational psychology, organizational and consulting tasks that involve personality assessment are commonly performed by counseling psychologists as well, with their traditional expertise in vocational psychology, developing human strengths, and the use of assessment tools. Personality measures commonly used in career counseling are often used for employee selection as well, as they may be better suited to predicting future employee performance than measures designed to detect psychopathology. Meta-analytic studies conducted in the 1990s demonstrated that personality measures had incremental validity in predicting job performance, and since that time, both research into and use of personality assessment for employee selection has increased exponentially (Rothstein & Goffin, 2006). Personality assessments based on the five-factor model have become increasingly popular for this purpose

In addition to employee selection purposes, personality testing can be useful in organizational consulting. For example, personality assessments are frequently used by counselors consulting in organizational settings to help select members for teams or to engage in team relationship development and team-building activities. Both the MBTI and Big-Five-based personality measures have been used extensively in this way, and also investigated in research (Kuipers, Higgs, Tolkacheva, & deWitte, 2009; Peeters, Rutte, van Tuijl, & Reymen, 2006).

IMPROVING OUTCOMES IN COUNSELING AND PSYCHOTHERAPY

A defining feature of assessment use for counselors and counseling psychologists is the use of test

results in counseling to stimulate client exploration and empower clients to make their own decisions, in contrast to a more clinical focus on helping the practitioner, not the client, to understand and conceptualize cases (Campbell, 2000). Thus, counseling psychologists have strong interests in the utility of personality assessment as a therapeutic intervention in and of itself. Finn (1996) described a collaborative assessment approach using the MMPI–2 that illustrates how personality assessment may be used as an intervention. In this approach, assessment begins by eliciting the client's reason for agreeing to take the test. Client and counselor then work together to develop a personalized set of "Assessment Questions" that the counselor and client use to guide the assessment and interpretation process. During feedback, counselors present their impressions based on the test data as relevant to the client's questions, and clients are encouraged to verify, modify, or reject the findings, and ultimately to summarize what has been learned. A program of experimental research using this approach with the MMPI–2 in college counseling centers has demonstrated beneficial effects of this brief intervention on psychological distress, self-esteem, and client hope (Finn & Tonsager, 1992; Newman & Greenway, 1997). The use of psychological assessment as an intervention has also been shown to have a positive effect on client outcome in general (Poston & Hanson, 2010).

Campbell (2000) describes a number of potential uses of tests in counseling that are consistent with an emphasis on the client as a collaborator in the assessment process, including using personality assessments to: provide psychoeducation, identify client strengths, teach a decision-making or problem-solving process, empower clients to make their own choices, foster self-actualization, gain perspective through comparison with others, clarify goals, and facilitate both self-awareness and self-exploration. However, this is again a case where the "specialness of the specialty" should not be overstated, as counseling psychologists routinely use personality assessments in counseling for case conceptualization, for psychodiagnosis or to select an area of focus, for forensic decision making (e.g., child custody, social security, or court-ordered evaluations), to predict outcome of counseling, to tailor or recommend interventions, and to track the outcome of counseling over time, all uses perhaps more traditionally associated with personality assessment by clinical as well as counseling psychologists.

IN RESEARCH

Counseling psychologists are also, of course, involved in using personality assessment in research. Counseling psychology journals have published research on personality assessment development and validation (e.g., Gough, 1969); adapting personality assessment measures for counseling settings (e.g., Duckworth, 1990), and personality in multicultural contexts (e.g., Ponterotto, 2010). Counseling psychologists continue to revitalize the field of personality assessment, evaluating and revising current measures to ensure they remain relevant, developing new measures, and applying personality assessment tools to questions and problems of interest to counseling psychologists.

HOW COUNSELORS ASSESS PERSONALITY

Both in research and in practice, personality assessments are used in diverse ways by counseling psychologists, so it is little surprise that the assessment tools counseling psychologists use are also diverse. Counseling psychologists use a wide variety of tools and procedures to assess personality, including assessment interviews, objective and projective psychological tests, qualitative procedures, and collateral data. In this section, we briefly describe and review the tools counseling psychologists most commonly use.

THE ASSESSMENT INTERVIEW

An initial clinical, assessment, diagnostic or intake interview is the single most commonly used assessment tool for counseling (May & Scott, 1991) and clinical psychologists (Watkins, Campbell, Nieberding, & Hallmark, 1995) alike, perhaps because it is the one assessment method readily available to all therapists without additional cost in time or money

beyond that already invested in the therapy hour (Cormier, Nurius, & Osborn, 2009). In addition, it is increasingly becoming the only assessment technique for which clients' insurers will pay (Eisman et al., 2000).

Despite its ubiquity, the unstructured clinical interview has long been the subject of substantial controversy as being of poor validity compared with objective tests (Sawyer, 1966); poorly defined and specified (Matarazzo, 1978); and difficult to effectively teach, research, or evaluate psychometrically (Smelson, Kordon, & Rudolph, 1997). One major problem in evaluating the utility of the clinical interview lies in the widely diverse theoretical orientations, goals, questions and other clinician behaviors that may be comprised in a "clinical interview"; these differences in counselor and client behavior from interview to interview may lead to unreliable data collection. Because of these disadvantages, some authors have advocated for replacing the unstructured clinical interviews with structured interviews that have demonstrated reliability (e.g., Zimmerman, 2003). However, others have suggested that structured clinical interviews are best used as an adjunct to more typical clinical interviewing, to answer specific questions about a client (Rogers, 2001).

The clinical interview does offer several advantages over other assessment tools, such as the opportunity to develop rapport with the client and a relatively ambiguous, unstructured situation that may offer greater opportunity for clients to demonstrate various aspects of personality untapped by an assessment battery (Groth-Marnat, 2009). The opportunity for the client to take an active, collaborative role in the assessment process is another benefit to the clinical interview that may be of special importance to counseling psychologists. Counseling psychology training programs appear to value the clinical interview as a major assessment tool; one study found that a course in clinical interviewing was required in nearly 70% of Counseling of Counseling Psychology Training Programs (CCPTP) member programs. Only coursework in vocational assessment was more commonly required (May & Scott, 1991).

Research into the clinical interview in counseling psychology has tended to focus on how attributes of counselors and their match to the wishes of clients affect the initial interview and outcome of counseling (e.g., Duckro & George, 1979; Hubble & Gelso, 1978), generally finding that such variables are of less importance to client satisfaction or outcome than had been hypothesized. Counseling researchers have also been interested in multicultural and bias issues in clinical interviewing, such as whether counselor–client match in terms of gender or ethnicity is important for initial rapport or eventual counseling outcome (Sue, 1988). Counseling psychologists have also been productive as teachers and textbook writers in clinical interviewing. Basic assessment and interviewing texts written by counseling psychologists are diverse in terms of theoretical perspectives; psychodynamic (Hill, 2009), cognitive–behavioral (Cormier, Nurius, & Osborn, 2009), and client-centered approaches (Daniels & Ivey, 2007) are all available. Regardless of theoretical orientation, however, an emphasis on client exploration and self-direction—in contrast to a medical or psychopathological model for case conceptualization—is an underlying theme in these texts that is highly consonant with traditional counseling psychology values. In addition, most counseling texts on clinical interviewing emphasize the importance of multimodal assessment and the use of other assessment tools beyond the clinical interview to gain an accurate picture of the client and facilitate collaborative treatment planning. These other commonly used tools in counseling are considered next.

OBJECTIVE MEASURES

Counseling psychologists, like other psychologists, commonly use objective personality tests, and objective tests have historically received a stronger emphasis in counseling than in clinical psychology, at least when compared with projective techniques (Watkins, Campbell, Hollifield, et al., 1989). Most counseling psychology training programs require a course in objective personality assessment, along with required coursework in clinical interviewing, intelligence and vocational assessment; only a minority of programs require training in other forms of personality assessment, such as projective testing (May & Scott, 1991; see also Chapter 11 in this

volume for additional information on the use of some objective measures of personality, although with a greater emphasis upon the assessment of psychopathology).

Several objective personality tests are commonly used by counselors (and others are strongly identified with the specialty), but most prominent for counselors as well as clinical psychologists may be the MMPI (or MMPI–2), which is the second most commonly used assessment procedure for both specialties (Watkins & Campbell, 1989; Watkins et al., 1995). The MMPI owes a debt, if not to counseling psychology, then at least to vocational psychology. It was originally developed in the late 1930s by Starke Hathaway, a clinical psychologist, and J. Charnley McKinley, a neuropsychiatrist, who borrowed from E. K. Strong's Vocational Interest Blank (Strong, 1935) the test construction method of identifying test items empirically based on how they were answered by contrasting groups. Hoping to differentiate individuals with schizophrenia and other specific psychiatric diagnoses from one another and from normal individuals, they applied this method to a pool of items taken from a variety of sources; although ultimately the individual scales were not able to provide definitive psychiatric diagnosis, the test was nevertheless rapidly adopted. Revised in 1989 to make the normal reference group more representative of the U.S. population as a whole and to eliminate dated or offensive items, among other changes, the MMPI–2 now includes 567 items, 10 clinical scales, a variety of validity and content scales, and the controversial Restructured Clinical (RC) scales (Tellegen et al., 2003). The RC scales represented an effort to reduce problems with covariation in the original clinical scales, but they have been criticized for their use of factor analysis (rather than empirical keying) in scale construction and for the lack of evidence that these scales are adequately similar to the original clinical scales which they are intended to replace[2] (Butcher, Hamilton, Rouse, & Cumella, 2006; Nichols, 2006).

It remains the most widely used personality assessment in the world (Nichols, 2001), but for counseling psychologists, one key problem with the MMPI-2 is the issue of how to interpret scales designed to measure psychopathology for normal individuals with problems in living. Duckworth and Anderson (1995) have provided an excellent text designed to respond to this challenge. In particular, they provide a detailed consideration of the meaning of moderate scale elevations for clients without marked psychopathology. They contrast the approach of interpreting moderate scale elevations as if they reflect attenuated psychopathology against a more novel approach (following the work of Kunce & Anderson, 1984) that posits personality traits underlying each clinical scale that may be either positive or negative. For example, a moderately elevated scale 7 (psychasthenia) in a normal individual may indicate a tendency to be well organized and methodical when functioning well, but obsessive or ritualistic when under stress. They also provide a framework to help counselors decide how to interpret moderate scale elevations, and to distinguish for particular clients whether moderate elevations indicate attenuated psychopathology, positive personality characteristics, or transient stress-elicited symptoms.

The California Personality Inventory (CPI), like the MMPI–2, is a (mostly) empirically keyed personality test with a rich history and tradition. Unlike the MMPI–2, however, it is specifically designed to assess normal personality, and thus is especially appropriate for counseling psychologists. Another attractive aspect of the CPI to counseling psychologists may be its pragmatism: The goal of the CPI is explicitly *not* to identify psychometrically "pure" or internally consistent traits, but instead to help construct a true-to-life picture of examinees, such that their behavior can be accurately predicted and they can be differentiated from others in terms of the perspective of those who know them well. Scales aim to represent "folk concepts," everyday ideas that ordinary people in all cultures use to describe others' personalities, such as responsibility or tolerance. Reflective of its goals, the CPI also has a unusual approach to psychometrics; in particular, the CPI is

[2]See the special issue on the restructured clinical scales in the 2006 *Journal of Personality Assessment* (Volume 87, No. 2) for a full range of critiques and rejoinders on this issue.

conceived of as an "open system" that can be easily modified to improve the predictive utility of the test; instead of providing internally consistent, minimally intercorrelated scales, the CPI seeks to allow scales to correlate just as certain traits (such as social dominance and sociability) intercorrelate in everyday life (Gough, 2000). Like the MMPI–2, the CPI has amassed a substantial research literature that enhances counselors' ability to interpret both individual scales and profile configurations (McAllister, 1996). Cross-cultural research has also been a major focus of inquiry with the CPI and the measure has been translated into more than 40 languages.

The Myers–Briggs Type Indicator (MBTI) may be the most commonly used theoretically or rationally derived objective personality assessment in counseling; one review described it as "the most widely used personality instrument for nonpsychiatric populations" (Murray, 1990, p. 1187). In particular, the MBTI has frequently been used in conjunction with the SII in career counseling, and resources are available to help counselors integrate the two measures (e.g., Hammer & Kummerow, 1996). The MBTI is also frequently used in organizational development settings for team building and managerial decision making (Moore, 1987). The MBTI is based on Jung's (1921/1971) theory of psychological types, which posits introversion–extraversion as a key dimension along which people vary. The theory also suggests that people may be further differentiated by the psychological functions they use in daily life, such as their preferences for judging versus perceiving as well as the way they perform judging and perceiving tasks (whether by thinking or feeling, in the case of judging, or sensing vs. intuiting, in the case of perceiving). The MBTI places examinees into one of 16 types, each signified by a four-letter code that references the examinee's preferred pole on each MBTI dimension (e.g., ENFP, meaning that the examinee preferred extraversion over introversion, intuition over sensing, feeling over thinking, and perceiving over judging).

The MBTI has been widely used and researched; however, it has also been criticized, particularly because of the conversion of examinees' scores on continuous measures of the four personality dimensions to dichotomous type categories that may obscure variations among people within types. There is little evidence that scores on the four dimensions are bimodally distributed, and individuals who score at the 5th and 35th percentile of, for example, the thinking–feeling scale are treated identically in terms of assigning their type, leading to questions about the predictive validity of types versus continuous scale scores (Healy, 1989, 2000). The MBTI has also been criticized as a managerial development tool because of the potential for types to be misused in organizations in a discriminatory or simplistic way (Healy, 1989; Michael, 2003).

In addition to empirically and theoretically derived objective personality tests, counseling psychologists have also made frequent use of factor-analytically derived tests, such as the 16PF and the Neuroticism-Extroversion-Openness Personality Inventory (NEO PI–R; Costa & McCrae, 1992a). Both these measures assess normal personality rather than psychopathology and are widely used by counseling psychologists in organizational, college counseling, and clinical settings (McCrae & Costa, 1991; Schuerger, 2000). Both also are structured hierarchically based on a five-factor model of personality, although these core factors are given different names in each instrument; whereas the NEO PI–R assesses neuroticism, extraversion, openness to experience, conscientiousness, and agreeableness, the 16PF instead assesses the similar constructs of anxiety, extraversion, independence, self control, and tough-mindedness, respectively. In addition to the assessment of the five factors, both tests also assess numerous underlying constructs, such as warmth, liveliness, and social boldness on the 16PF, or gregariousness, assertiveness, and excitement seeking on the NEO PI–R. A unique aspect of the 16PF is the variety of available forms and interpretive material for special purposes; adolescent, child, clinical, and low-literacy forms of the test are available as well as supplemental material and special interpretive reports for couples, career development, and personnel and clinical evaluation. The NEO PI–R is distinguished by the availability of both self and observer rating forms, so that self-reported personality can be compared against the ratings of others who know the examinee well.

PROJECTIVE MEASURES

Projective personality assessment techniques (now often referred to as performance-based personality measures; see Chapter 10, this volume) are distinguished from objective assessments by the unstructured quality of the tasks examinees are asked to complete; instead of selecting from a limited array of choices (e.g., true/false) for how to respond to a test item, examinees undergoing projective assessment have a nearly unlimited variety of possible responses available to them. The general hypothesis behind projective techniques is that the examinee will perceive and respond to test material in such a way as to "project" their (often unconscious) personality characteristics onto the task (Anastasi & Urbina, 1997). Because the way in which responses are interpreted is unclear or may be disguised from the examinee, projective techniques are often used to assess psychopathology, deviant behavior or other factors examinees might seek to avoid disclosing. The emphasis on unconscious processing and psychopathology in projective techniques may be one reason these tests are more strongly aligned with clinical psychology and may help explain counseling psychology's ambivalence about them.

Nevertheless, counseling psychologists commonly use projective personality assessment techniques. About 45% of counseling psychologists in Division 17 reported using projectives in practice (Fitzgerald & Osipow, 1986). Sentence completion blanks, the TAT (Murray, 1971) the Bender–Gestalt (Pascal & Suttell, 1951), and projective drawing techniques (Rabin, 1986) are all represented in the top ten most commonly used assessment techniques by counseling psychologists (see Table 24.1). Compared with clinical psychologists, however, counseling psychologists may use projective techniques less commonly; the Rorschach, fifth most commonly used by clinical psychologists and perhaps the quintessential projective test, does not appear on the top 10 list for counseling psychologists at all.

Counseling psychology training programs reflect the ambivalence of the field about projectives. A minority of programs require coursework in projectives, and a larger minority offer training through electives; whereas nearly all training directors believe training in objective personality assessment is of major or substantial importance, fewer than 25% say the same about training in projective assessment. Nevertheless, most training directors believed their students should learn to administer and interpret projective tests, albeit with less emphasis than in clinical psychology training programs (Watkins, Campbell, & Manus, 1990).

Some authors have argued that projective techniques deserve a more prominent position in counseling psychology training and practice (Clark, 1995; Watkins, Campbell, Hollifield, et al., 1989), particularly as qualitative tools that can be used to develop hypotheses about a client or to enhance the therapeutic alliance or other outcomes in psychotherapy, rather than as norm-referenced psychological tests, per se. In this respect, the reservations counseling psychologists may hold about projective assessments may be more a function of the measures usually identified with the term than with resistance to the use of open-ended, projective, qualitative tools in counseling. Vocational card sorts, for example, are not usually identified as projective personality assessments, and yet, as Goldman (1983) noted, they require examinees to project idiosyncratic personal material onto the classification of occupational titles in a fashion highly similar to other projective tests.

OTHER METHODS OF PERSONALITY ASSESSMENT

Counseling psychologists use other methods of personality assessment that do not fit neatly into "projective" or "objective" personality testing categories. Goldman (1990) provided a useful overview of what he described as "qualitative" assessment procedures, which he defined as nontraditional assessment procedures that involve several key elements, including: an active role for the client in collecting and interpreting the assessment data, a holistic and integrative viewpoint on the individual, an emphasis on learning about oneself as a developmental change process, and a blurring of the distinction between assessment and counseling or intervention. Examples of qualitative assessment included the Vocational Card Sort (VCS; Dewey, 1974) just described;

Lifeline methods, where clients are asked to draw a line representing the course of their lives and place life events into various categories of meaning; situational assessments such as work samples or the in-basket test, and observations of behavior, such as job shadowing experiences used by clients in career counseling. Such methods do not always lend themselves well to traditional psychometric evaluation, although researchers have attempted to ascertain and improve the reliability and validity of many of these procedures, such as the in-basket test (e.g., Brannick, Michaels, & Baker, 1989). Instead, qualitative assessment procedures may function more as interventions or hypothesis-generating methods, where later experiences in counseling may alter the meaning of assessment results for either or both counselor or examinee.

MULTICULTURAL ISSUES IN PERSONALITY ASSESSMENT IN COUNSELING

Counseling psychologists make efforts to be culturally competent in using personality assessment tools and processes with diverse clients. In this section, we begin by defining culture and providing definitions of culture-related concepts that may influence personality and personality assessment. Next, we briefly review efforts at creating more multiculturally competent personality assessment tools, including biases and limitations that affect these tools and strategies used to develop more culturally competent assessment.

Culture is a broad concept that we define here as a dynamic set of systems of meanings which are learned and shared by a group of people transmitted across generations for the purpose of human adjustment and development (Dana, 2000). External referents of culture include artifacts, roles, and institutions; internal referents include attitudes, values, beliefs, expectations, epistemologies, and consciousness (Ridley, Li, & Hill, 1998). Culture has an effect on psychological outcomes in a myriad of ways. Culture influences the social construction of reality, which, in turn, affects self-concept, emotion regulation, and personality. The importance of understanding how culture influences psychology in general and personality in particular is a major challenge to counseling psychologists, to ensure that traits are interpreted within a cultural context and to prevent misdiagnoses and misuse of assessment tools (Ridley et al., 1998).

Culture provides a set of contextual variables that moderate behavior and shape personality (Lonner & Adamopoulos, 1997). Examples of culture-related variables that can affect how personality develops or is expressed on an assessment tool include: individualism versus collectivism; race, ethnicity, and perceived discrimination; and acculturation and acculturative stress. Cultural values such as individualism and collectivism can variably shape the meaning of personality itself from one culture to another. Individual differences in the extent to which one identifies with one's cultural group or groups can also contribute to between-group and within-group differences in personality. Conceptions of self and personality can also be affected by how one has negotiated differences between cultural values in cases where two or more cultures intersect, or by other elements of the sociocultural context, such as the presence of oppression or discrimination (Church et al., 2006; Matsumoto, 2006).

CULTURE-RELATED CONCEPTS THAT INFLUENCE PERSONALITY

The validity of psychological assessment is increased when it occurs within a cultural context (Ridley et al., 1998). Adjusting for the potential moderating role of various culture-related variables can increase the validity of the personality assessment process. A variety of culture-related concepts have been identified that can affect assessment results, such as acculturation, acculturative stress, and ethnic identity, although these identified factors are not the only variables to consider in exploring how clients' cultural backgrounds may have an effect on assessment results.

Acculturation refers to multilevel changes that occur in one or more cultural groups and individuals that come into continuous, firsthand contact with one another. The interaction and exchange of information between groups or individual members of groups lead to subsequent changes in the original cultural patterns of either or both groups and

individual group members (Cuéllar, 2000). Acculturation is often measured using Berry's (1995) model, which delineates four distinct stages of acculturation: traditional, bicultural, assimilated, and marginal. These categories help to differentiate the varying levels at which individuals negotiate the retention of their heritage (or previous) culture and adoption of values held by the majority (or contact) culture. The first stage, traditional, refers to the retention of most or all of one's original culture, with little to no acceptance of the majority culture. Biculturality characterizes those who are familiar with both their heritage culture and the majority culture, and who have the flexibility to equally engage in both worlds. Assimilation refers to individuals who have primarily adopted and internalized the values, beliefs, and behaviors of the majority culture. Finally, marginal status refers to having acquired neither a comfortable acceptance of one's heritage nor majority culture and values. The process of acculturation affects six major areas of psychological functioning: language, cognitive styles, personality, identity, attitudes, and acculturative stress (Berry, 1995), which, in turn, can affect personality assessment results either directly (i.e., by changes in personality by acculturation stage) or indirectly (e.g., because attitudes toward the assessment situation differ by acculturation stage and therefore affect how clients respond, even if underlying personality constructs remain the same across levels of acculturation). *Acculturative stress,* a related but distinct concept, refers to the tension between the pressures to acculturate to the dominant society and the pull toward one's ethnic group. This stress can lead to psychological maladjustment and can result in changes to personality and psychological outcomes.

The Acculturation Rating Scale for Mexican Americans (ARSMA), Forms I and II (Cuéllar, Arnold, & Maldonado, 1995), is one of the most common assessment tools used to assess level of acculturation. Modifications of the ARSMA also exist for African Americans, American Indians, Alaska Natives, and Asian Americans (Cuéllar, 2000). ARSMA–II (Cuéllar, 2000) provides both linear and orthogonal measures of acculturation; orthogonal scales allow for the measurement of bicultural orientations that are not differentiated when using linear measures. Acculturation as assessed by the ARSMA or one of its modifications can be used as a moderator variable in personality assessments to provide multidimensional information on culture-specific attitudes and values such as individualism and collectivism, language proficiency and preference, and socioeconomic status. Including measures of acculturation affords the detection of both within-group and between-group differences when assessing personality. Research has demonstrated that acculturation level can have an effect on personality and personality dysfunctions, with some research suggesting that clinical measures of personality may be more sensitive to acculturation level than other personality traits. For example, three MMPI clinical scales (scales 4, 6, and 8) are posited to be more sensitive to the moderating effect of acculturation (Cuéllar, 2000).

Ethnic identity serves as another culture-related variable that may moderate personality assessment results. Ethnic identity is part of one's self-concept that refers to the acquisition and retention of cultural characteristics that are incorporated into one's self-concept, and which develop in the context of one's membership to a minority ethnic group within the larger society (Phinney, 1992). Multiple models of ethnic identity exist and most propose both categorical and continuous ways of measuring levels of ethnic identity development. One of the more widely applied models is Phinney's process model of ethnic identity formation, which is theoretically based on Erickson's model of ego development (Phinney, 1992). Researchers have posited that higher levels of ethnic identity are related to positive psychological adjustment, high self-esteem, self-confidence, and a sense of purpose in life (Farver, Narang, & Bhadha, 2002).Some studies have found that ethnic identity development is related to personality traits, such as those assessed by the five-factor model, or that ethnic identity may moderate the relations between personality traits and other outcomes (Roysircar-Sodowsky & Maestas, 2000). For example, the white racial identity development stage of pseudoindependence, which is characterized by distorting incoming information about race to be consistent with a "liberal" worldview, is a

significant negative predictor of openness to experience, whereas the autonomy stage, characterized by flexible interaction with and a complex understanding of racial stimuli, is a positive predictor (Silvestri & Richardson, 2001). Hurlic (2009) found that ethnic identity development moderated the relationship between five-factor model personality traits and attitudes toward affirmative action policies.

TOWARD MULTICULTURALLY COMPETENT PERSONALITY ASSESSMENT

A major challenge for counseling psychologists who use personality assessments is to ensure the multicultural competence of tests. The potential for cultural bias in personality testing, particularly as it is used for psychodiagnosis, is well known (Malgady, 1996; Marsella & Leong, 1995). Tests may be culturally biased because of inappropriate item content, inappropriate standardization samples, examiner or test user biases, or cultural differences in approach to the testing situation (Nichols, Padilla, & Gomez-Maqueo, 2000). Another potential area of concern is that the constructs measured by a personality test—or the construct of personality itself—as developed in a particular culture are specific to that culture, or emic, and cannot be generalized to members of other cultural groups even when other sources of bias are eliminated.

Counseling psychologists have a responsibility to ensure that personality assessments are used in culturally appropriate ways, and where necessary, to use culture-specific instruments that more accurately assess personality of test takers from various cultures. Either etic or emic approaches can be used to create culturally appropriate personality measures or to evaluate the cross-cultural applicability of tests. Tests of various types of equivalence are another way in which the performance of personality assessments across cultures can be evaluated.

Etic approaches use universal, culture-general concepts to assess personality across cultures, whereas emic approaches conceptualize personality from the internal perspective of each specific culture under consideration. Personality assessments have been developed using each approach; for example, the CPI, widely translated and used in multiple cultures, takes an etic approach by identifying an "open system" of culturally universal personality constructs, whereas Ko's Mental Health Questionnaire (Ko, Yang, Cheng, & Li, 1975) was developed indigenously in China from an emic perspective. Combined approaches have also been used to develop culture-specific personality measures, such as the Chinese Personality Assessment Inventory (Cheung et al., 1996), which includes both etic or universal subscales and subscales derived using an emic approach.

Tests of equivalence can also be used to assess how well a given assessment tool generalizes across cultures and can be interpreted across cultures. Key types of test equivalence include linguistic or translational equivalence, conceptual equivalence, and metric equivalence (Nichols et al., 2000). Linguistic equivalence refers to how well test validity is preserved after items are translated into different languages, and is generally accomplished through back translation, where a test is first translated from its initial language (e.g., English) into another language (e.g., Chinese), and then translated again from Chinese back into English. If the back-translated version of the test is not meaningfully or psychometrically different from the original version, linguistic equivalence has been achieved.

Conceptual equivalence refers to whether the meaning and subjective experience of the assessed constructs are equivalent cross-culturally, and that the meaning ascribed to a construct is comparable across cultures. As with construct validity, conceptual equivalence is difficult to establish conclusively, but the use of a multitrait–multimethod approach in test development with a variety of cultures as well as the use of statistical techniques such as confirmatory factor analysis, may assist in establishing conceptual equivalence.

Metric equivalence is established when a scale demonstrates similar psychometric properties across cultures. Typically, this equivalence can be established by analyzing the rate of item endorsement across samples, with a difference of less than 25% suggesting adequate equivalence. Significant differences in rates of endorsement of items across cultures suggest there may be translational and/or conceptual nonequivalence (Nichols et al., 2000).

Although much remains unknown about how best to measure personality across and within cultures, research does exist that permits better integration of cultural factors into the personality assessment process. Additional research is needed, along with improvements in practice.

RECOMMENDATIONS FOR RESEARCH AND COUNSELING PRACTICE

We consider next recommendations for change in research and counseling practice in personality assessment, both with regards to multicultural concerns as well as personality assessment in general.

A Need for Advocacy

Perhaps the most important recommendation we could make to counseling psychologists with respect to personality assessment falls not under the rubric of research or practice, but advocacy. The future of personality assessment in counseling psychology, and for psychology as a whole, appears threatened. A major report by the APA's Psychological Assessment Work Group (PAWG; Eisman et al., 2000) concluded that psychological assessment services are under assault from organized health care delivery systems, managed care organizations, and health care payers. Psychologists described difficulties with preauthorization and reimbursement from third-party payers for psychological testing even when it was strongly indicated. Many insurance preauthorization decisions appeared to the PAWG to be driven by economics rather than by clinical concerns, with some national managed care organizations paying less per hour for psychological assessment than for individual therapy. In addition the PAWG reported that most health care companies refused to reimburse psychologists for assessment done by appropriately trained and supervised students, interns, or unlicensed postdoctoral psychologists, creating a barrier to training for future psychologists.

In the 10 years since this report was published, there is some evidence of modest change in the situation that the PAWG documented, but little sign of sweeping reform. At the federal level, some recent changes in Medicare have benefitted the practice of assessment, such as allowing psychologists (not just physicians) to supervise technicians in doing testing. Some survey data suggests that counseling psychologists who provide specialty services such as forensic, neuropsychological, or medical evaluations are experience growing rather than declining opportunities for assessment (Rich, 2007). However, in general, both managed care and Medicare reimbursement for psychological testing continues to experience cuts, and 47% of psychologists in one recent survey indicated that the market for testing services was shrinking (Rich, 2007).

Because of threats like these, the PAWG concluded (Eisman et al., 2000) that as a profession, psychology must respond with advocacy and a credible explanation of the value and usefulness of assessment if it is to survive as a covered health care service. Indeed there is abundant evidence that psychological test validity, broadly speaking, is strong and compelling, comparable with medical test validity; that assessment procedures offer important incremental validity above that provided by a clinical interview alone; and that psychological assessment is useful for a variety of purposes in and beyond health care settings (Kubiszyn et al., 2000; Meyer et al., 2001). In addition, psychological assessment is one of the few professional activities uniquely associated with psychology and in which psychologists can claim special expertise relative to other mental health professional service fields.

We agree with the PAWG that counseling psychologists should respond vigorously to these external barriers to wise and responsible use of personality assessments. We suggest that counseling psychologists should participate in and lead efforts to advocate for the continued use and reimbursement of psychological testing, at all levels, from local clinic decision making to federal policy. Counseling psychologists can take roles in political advocacy, in research establishing and disseminating the validity and utility of personality assessment procedures, and in providing assessment services that persuade test users and local decision makers of the value of the service. For example, counseling psychologists who use personality assessments should make efforts to use best practices; in particular, counseling psychologists should make efforts to individualize test batteries, interpretation reports, and feedback processes to suit individual clients and

the specific referral question (Brenner, 2003). Implementing consumer-focused personality assessment practices may enhance our ability to counteract negative stereotypes about the utility of personality assessment in the minds of test users and health care leaders.

Integrating Assessment With Counseling Interventions

Personality assessment can be a powerful intervention. A recent meta-analysis suggested that personality assessment with client feedback, when delivered in a collaborative, personally involving fashion, has beneficial effects on client outcome, comparable in size with those of other psychological interventions, such as substance abuse treatment (Poston & Hanson, 2010). However, despite the utility of personality assessment for this as well as other purposes, we suspect that the threats to its continued use come not only from outside the profession but also from counseling psychologists themselves. Several authors have documented a substantial decline in the number of counseling psychologists involved or interested in vocational assessment (Fitzgerald & Osipow, 1986; Goodyear et al., 2008). It appears this unfortunate decline may extend to personality assessment as well. Whereas about 45% of Society of Counseling Psychology ([SCP] i.e., APA Division 17) members reported personality diagnosis or assessment as a professional activity in 1985; by 2000, fewer than 20% endorsed this item. Among counseling psychologists who were not SCP members, about 37% engaged in personality diagnosis or assessment, suggesting that the decline may be even sharper among those more closely identified with the specialty (Goodyear et al., 2008). We do not know the reasons for this decline in test use by counseling psychologists—whether, for example, it is best explained by external barriers to test use or by internal barriers, such as lack of interest or faith in the utility of the measures—but we believe that both types of barriers should be addressed.

Some authors have suggested that vocational assessment could be reinvigorated by using qualitative assessment tools such as the VCS to better integrate assessment with counseling and more deeply engage both counselors and clients in the work of career exploration and self-understanding (Goldman, 1990; Slaney & MacKinnon-Slaney, 2000). Qualitative assessment methods foster a more active role for the client in a developmental process of self-discovery that may be more adaptable to clients from diverse and underrepresented backgrounds and to a variety of assessment settings. However—and, it is important to note, from the perspective of addressing internal barriers to effective assessment use—these methods may offer benefits to the counselor in terms of creating an assessment situation that counselors as well as clients experience as more intimate, involving, and creative than traditional personality assessment. Although we are aware of no empirical data to support the suggestion that qualitative assessment tools lead to improved counselor or client engagement with the assessment process, we view this hypothesis as worthy of further research. Although qualitative assessment methods are relatively well developed in vocational and personnel assessment, we know of few qualitative personality assessments, although several projective tests, as well as Q-sort personality measures (Block, 1961), could likely be adapted to fit the criteria of qualitative assessment. Development of qualitative personality assessments suited for counseling, thus, is another area that we suggest would be fruitful for research. In addition to research, counseling psychology is likely to benefit from the incorporation and integration of qualitative methods into both personality assessment training and practice.

Computer and Internet-Based Testing

In some respects, we find the difficulties faced by psychologists in promoting and using personality tests puzzling, because personality tests have an obvious and enduring appeal, as even a cursory browse through the Internet reveals. On social networking websites such as Facebook, applications that allows users to "test their personalities" are highly popular; for example, the Facebook application Quiztastic, which allows users to "become a contestant in one of thousands of quizzes and personality tests," boasts more than 300,000 active monthly users. Facebook applications purporting to offer test results based on psychological (or quasi-psychological) theories such as "The Sorting Hat

quiz (based on Briggs–Myers Personality Quiz)" or "What Enneagram Personality Type Are You?" are also popular. Internet dating websites also make prominent use of personality testing. Popular dating site OKCupid, for example, claims to use advanced mathematical modeling and (undisclosed) psychometric techniques to provide "superior personality analysis" to match users based on their responses to a variety of "personality tests" written by the site authors and by users.

Despite the evidence that many nonpsychologists are making a living providing "personality testing" for romance, recreation, or self-improvement online, psychologists have been appropriately cautious about providing personality assessment online. Personality assessment in computerized and Internet-based formats raises a variety of ethical, legal, clinical, and professional concerns, including problems with maintaining test security, psychometric concerns when translating paper-and-pencil tests to computer and Internet-based formats, difficulties with maintaining client confidentiality, problems with practice jurisdiction for psychologists who administer tests online, and ethical problems related to providing psychotherapy or assessment interpretation to a client you may never actually see. These ethical and professional concerns represent serious barriers to providing assessment services in online or computerized formats, and yet we believe that more involvement by counseling psychologists in computer-based and Internet-based personality assessment may be professionally and clinically beneficial, provided ethical and professional concerns can be addressed effectively.

Promoting Multiculturally Competent Personality Assessment

Despite increased attention to the issues involved, there continues to be a need for personality assessment procedures with demonstrated cross-cultural utility. In addition, existing measures that do effectively address cultural considerations need to be more widely disseminated and used in place of measures whose cross-cultural utility is untested or known to be biased. Dana (1996), for example, suggests the Holtzman Inkblot Test (Holtzman, 1975) as an alternative to the Exner Rorschach (Exner, 1969) that is superior both psychometrically and in terms of cross-cultural validity. In addition to using appropriate assessment tools, counseling psychologists should follow a multiculturally competent assessment process, such as the Multicultural Assessment Procedure (Ridley, Li, & Hill, 1998), which flexibly integrates cultural and individual information and minimizes the likelihood of bias.

Despite threats, personality assessment remains a major activity of counseling psychologists. Assessments can be used as intervention tools to provide brief but potent benefits to clients (Poston & Hanson, 2010), or as decision-making tools to make reliable and valid inferences about clients dispositions and future behaviors (Meyer et al., 2001). When misused, they can also cause harm both to clients and to counseling psychology as a profession, a risk that may be especially acute with clients from traditionally oppressed or cultural minority groups. We believe counseling psychologists should continue to advocate for the appropriate use of personality assessment; to research, disseminate, and use best practices in assessment; and work to invigorate our own practice of assessment as a dynamic, involving, growth process for our clients. Personality assessment is a basic element of our profession, and one with which we can promote social change and client well-being and self-actualization.

References

American Psychological Association. (n.d.). *Archival description of counseling psychology*. Retrieved from http://www.apa.org/crsppp/counseling.html

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Beck, A. T., & Steer, R. A. (1987). *Beck Depression Inventory manual*. San Antonio, TX: Psychological Corporation.

Berry, J. W. (1995). Psychology of acculturation. In N. R. Goldberger & J. B. Veroff (Eds.), *The culture and psychology reader* (pp. 457–488). New York, NY: New York University Press.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C Thomas. doi:10.1037/13141-000

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*, 957–963. doi:10.1037/0021-9010.74.6.957

Brenner, E. (2003). Consumer-focused psychological assessment. *Professional Psychology: Research and Practice, 34*, 240–247. doi:10.1037/0735-7028.34.3.240

Brown, S. D., & Lent, R. W. (2000). *Handbook of counseling psychology* (3rd ed.). New York, NY: Wiley.

Brown, S. D., & Lent, R. W. (2008). *Handbook of counseling psychology* (4th ed.). New York, NY: Wiley.

Brown, S. D., & Ryan Krane, N. E. (2000). Four (or five) sessions and a cloud of dust: Old assumptions and new observations about career counseling. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 740–766). New York, NY: Wiley.

Buckworth, J., Granello, D. H., & Belmore, J. (2002). Incorporating personality assessments into counseling to help college students adopt and maintain exercise behaviors. *Journal of College Counseling, 5*, 15–25. doi:10.1002/j.2161-1882.2002.tb00203.x

Butcher, J. N., Hamilton, C. K., Rouse, S. V., & Cumella, E. J. (2006). The deconstruction of the Hy scale of MMPI–2: Failure of RC3 in measuring somatic symptom expression. *Journal of Personality Assessment, 87*, 186–192. doi:10.1207/s15327752jpa8702_08

Campbell, V. L. (2000). A framework for using tests in counseling. In C. E. Watkins Jr. & V. L. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 3–11). Mahwah, NJ: Erlbaum.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *The handbook for the Sixteen Personality Factor (16PF) Questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

Cheung, F. M., Leung, K., Fan, R. M., Song, W., Zhang, J., & Zhang, J. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology, 27*, 181–199. doi:10.1177/0022022196272003

Church, A. T., Katigbak, M. S., Del Prado, A. M., Ortiz, F. A., Mastor, K. A., Harumi, Y., Cabrera, H. F. (2006). Implicit theories and self-perceptions of traitedness across cultures. *Journal of Cross-Cultural Psychology*, *37*, 694–716. doi:10.1177/0022022106292078

Clark, A. J. (1995). Projective techniques in the counseling process. *Journal of Counseling and Development, 73*, 311–316. doi:10.1002/j.1556-6676.1995.tb01754.x

Cormier, S., Nurius, P. S., & Osborn, C. J. (2009). *Interviewing and change strategies for helpers*. Belmont, CA: Brooks/Cole.

Costa, P. T., Jr., & McCrae, R. R. (1992a). *NEO Personality Inventory—Revised (NEO PI–R) and NEO Five-Factor Inventory (NEO FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992b). Trait psychology comes of age. In J. J. Berman & T. B. Sondregger (Eds.), *Psychology and aging: Nebraska symposium on motivation* (pp. 169–204). Lincoln: University of Nebraska Press.

Cuéllar, I. (2000). Acculturation as a moderator of personality and psychological assessment. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 113–129). Mahwah, NJ: Erlbaum.

Cuéllar, I., Arnold, B., & Maldonado, R. (1995). Acculturation rating scale for Mexican Americans-II: A revision of the original ARSMA scale. *Hispanic Journal of Behavioral Sciences, 17*, 275–304. doi:10.1177/07399863950173001

Dana, R. H. (1996). Culturally competent assessment practice in the United States. *Journal of Personality Assessment, 66*, 472–487. doi:10.1207/s15327752jpa6603_2

Dana, R. H. (2000). An assessment-intervention model for research and practice with multicultural populations. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 5–16). Mahwah, NJ: Erlbaum.

Daniels, T., & Ivey, A. (2007). *Microcounseling: Making skills training work in a multicultural world.* Springfield, IL: Charles C Thomas.

Dewey, C. R. (1974). Exploring interests: A non-sexist method. *Personnel and Guidance Journal, 52*, 311–315. doi:10.1002/j.2164-4918.1974.tb04032.x

Duckro, P. N., & George, C. E. (1979). Effects of failure to meet client preference in a counseling interview analogue. *Journal of Counseling Psychology, 26*, 9–14. doi:10.1037/0022-0167.26.1.9

Duckworth, J. (1990). The counseling approach to the use of testing. *The Counseling Psychologist, 18*, 198–204. doi:10.1177/0011000090182002

Duckworth, J. C., & Anderson, W. P. (1995). *MMPI and MMPI–2: Interpretation manual for counselors and clinicians* (4th ed.). Levittown, PA: Accelerated Development.

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. (2000). Problems and limitations in the use of psychological assessment in the contemporary health care delivery system. *Professional Psychology: Research and Practice*, *31*, 131–140. doi:10.1037/0735-7028.31.2.131

Exner, J. E. (1969). *The Rorschach systems*. New York, NY: Grune & Stratton.

Farver, J. M., Narang, S. K., & Bhadha, B. R. (2002). East meets west: Ethnic identity, acculturation, and conflict in Asian Indian families. *Journal of Family Psychology, 16*, 338–350. doi:10.1037/0893-3200.16.3.338

Fee, A. F., Elkins, G. R., & Boyd, L. (1982). Testing and counseling psychologists: Current practices and implications for training. *Journal of*

Personality Assessment, 46, 116–118. doi:10.1207/s15327752jpa4602_1

Finn, S. E. (1996). *Using the MMPI–2 as a therapeutic intervention*. Minneapolis: University of Minnesota Press.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment, 4*, 278–287. doi:10.1037/1040-3590.4.3.278

Fitzgerald, L. F., & Osipow, S. H. (1986). An occupational analysis of counseling psychology: How special is the specialty? *American Psychologist, 41*, 535–544. doi:10.1037/0003-066X.41.5.535

Gallagher, J. J. (1953). MMPI changes concomitant with client-centered therapy. *Journal of Consulting Psychology*, *17*, 334–338. doi:10.1037/h0060815

Goldman, L. (1983). The Vocational Card Sort technique: A different view. *Measurement and Evaluation in Guidance, 16*, 107–109.

Goldman, L. (1990). Qualitative assessment. *The Counseling Psychologist, 18*, 205–213. doi:10.1177/0011000090182003

Goodyear, R. K., Murdock, N., Lichtenberg, J. W., McPherson, R., Koetting, K., & Petren, S. (2008). Stability and change in counseling psychologists' identities, roles, functions, and career satisfaction across 15 years. *The Counseling Psychologist, 36*, 220–249. doi:10.1177/0011000007309481

Gough, H. G. (1969). A leadership index on the California Psychological Inventory. *Journal of Counseling Psychology, 16*, 283–289. doi:10.1037/h0027717

Gough, H. G. (2000). The California Psychological Inventory. In C. E. Watkins Jr. & V. L. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 45–71). Mahwah, NJ: Erlbaum.

Groth-Marnat, G. (2009). *Handbook of psychological assessment*. Hoboken, NJ: Wiley.

Hammer, A. L., & Kummerow, J. M. (1996). *Strong and MBTI career development guide* (rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide*. Stanford, CA: Stanford University Press.

Hathaway, S. R., & McKinley, J. C. (1951). *The Minnesota Multiphasic Personality Inventory* (rev. ed.). New York, NY: Psychological Corporation.

Hathaway, S. R., & McKinley, J. C. (1991). *Minnesota Multiphasic Personality Inventory—2*. Minneapolis: University of Minnesota Press.

Healy, C. C. (1989). Negative: The MBTI: Not ready for routine use in counseling. *Journal of Counseling and Development, 67*, 487–488. doi:10.1002/j.1556-6676.1989.tb02125.x

Healy, C. C. (2000). Interpreting the Myers–Briggs Type Indicator to help clients in understanding their Strong Interest Inventory. *Journal of Career Development, 26*, 295–308. doi:10.1177/089484530002600405

Hill, C. E. (2009). *Helping skills: Facilitating exploration, insight, and action*. Washington, DC: American Psychological Association.

Holtzman, W. H. (1975). New developments in the HIT. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3, pp. 243–274). San Francisco: Jossey-Bass.

Hubble, M. A., & Gelso, C. J. (1978). Effect of counselor attire in an initial interview. *Journal of Counseling Psychology, 25*, 581–584. doi:10.1037/0022-0167.25.6.581

Hurlic, D. (2009). Diversity congruency within organizations: The relationship among emotional intelligence, personality structure, ethnic identity, organizational context and perceptions of organizational diversity (UMI No. 3350471). *Dissertation Abstracts International A,* 70(03).

Itzhar-Nabarro, Z., Silberschatz, G., & Curtis, J. T. (2009). The Adjective Check List as an outcome measure: Assessment of personality change in psychotherapy. *Psychotherapy Research, 19*, 707–717. doi:10.1080/10503300902988760

Jastek, S., & Wilkinson, G. S. (1984). *Wide Range Achievement Test—Revised (WRAT–R)*. Wilmington, DE: Jastak Associates.

Jung, C. G. (1971). Psychological types. (H. G. Baynes, Trans., rev. by R. F. C. Hull). *The collected works of C. G. Jung* (Vol. 6). Princeton, NJ: Princeton University Press. (Original work published 1921)

Ko, Y. H., Yang, K. S., Cheng, H. H., & Li, P. H. (1975). The relationship between school environment and college student mental health [in Chinese]. *Bulletin of the Institute of Ethnology*, *39*, 125–149.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice*, *31*, 119–130. doi:10.1037/0735-7028.31.2.119

Kuipers, B. S., Higgs, M. J., Tolkacheva, N. V., & deWitte, M. C. (2009). The influence of Myers–Briggs Type Indicator profiles on team development processes: An empirical study in the manufacturing industry. *Small Group Research, 40*, 436–464. doi:10.1177/1046496409333938

Kunce, J., & Anderson, W. (1984). Perspectives on uses of the MMPI in nonpsychiatric settings. In P. McReynolds & G. J. Chelune (Eds.), *Advances*

in psychological assessment (pp. 41–76). San Francisco: Jossey-Bass.

Larson, P. C., & Agresti, A. A. (1992). Counseling psychology and neuropsychology: An overview. *The Counseling Psychologist, 20*, 549–555. doi:10.1177/0011000092204001

Lonner, W. J., & Adamopoulos, J. (1997). Culture as antecedent to behavior. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds), *Handbook of cross-cultural psychology: Vol. 1. Theory and method* (2nd ed., pp. 43–84), Boston: Allyn and Bacon.

Lubin, B., Larsen, R. M., Matarazzo, J. M., & Seever, M. (1985). Psychological test usage patterns in five professional settings. *American Psychologist, 40*, 857–861. doi:10.1037/0003-066X.40.7.857

Malgady, R. G. (1996). The question of cultural bias in assessment and diagnosis of ethnic minority clients: Let's reject the null hypothesis. *Professional Psychology: Research and Practice, 27*, 73–77. doi:10.1037/0735-7028.27.1.73

Marsella, A. J., & Leong, F. T. L. (1995). Cross-cultural issues in personality and career assessment. *Journal of Career Assessment, 3*, 202–218. doi:10.1177/106907279500300207

Matarazzo, J. D. (1978). The interview: Its reliability and validity in psychiatric diagnosis. In B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A handbook* (pp. 47–96). New York, NY: Plenum Press. doi:10.1007/978-1-4684-2490-4_3

Matsumoto, D. (2006). Are cultural differences in emotion regulation mediated by personality traits? *Journal of Cross-Cultural Psychology, 37*, 421–437. doi:10.1177/0022022106288478

May, T. M., & Scott, K. J. (1991). Assessment in counseling psychology: Do we practice what we teach? *The Counseling Psychologist, 19*, 396–413. doi:10.1177/0011000091193009

McAllister, L. W. (1996). *A practical guide to CPI interpretation*. Palo Alto, CA: Consulting Psychologists Press.

McCrae, R. R., & Costa, P. T. (1991). The NEO Personality Inventory: Using the five-factor model in counseling. *Journal of Counseling and Development, 69*, 367–372. doi:10.1002/j.1556-6676.1991.tb01524.x

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Read, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Michael, J. (2003). Using the Myers–Briggs Type Indicator as a tool for leadership development? Apply with caution. *Journal of Leadership and Organizational Studies, 10*, 68–81. doi:10.1177/107179190301000106

Monsen, J., Odland, T., Faugli, A., Daae, E., & Eilertsen, D. (1995). Personality disorders: Changes and stability after intensive psychotherapy focusing on affect consciousness. *Psychotherapy Research, 5*, 33–48. doi:10.1080/10503309512331331126

Moore, T. (1987, March). Personality tests are back. *Fortune, 30,* pp. 74–78.

Murray, H. A. (1971). *Thematic Apperception Test manual*. Cambridge, MA: Harvard University Press.

Murray, J. B. (1990). Review of the Myers–Briggs Type Indicator. *Perceptual and Motor Skills, 70*, 1187–1202.

Myers, I. B., & McCauley, M. H. (1985). *Manual: A guide to the development and use of the Myers–Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Newman, M. L., & Greenway, P. (1997). Therapeutic effects of providing MMPI–2 test feedback to clients at a university counseling service: A collaborative approach. *Psychological Assessment, 9*, 122–131. doi:10.1037/1040-3590.9.2.122

Nichols, D. S. (2001). *Essentials of MMPI–2 assessment*. New York, NY: Wiley.

Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI–2 restructured clinical scales. *Journal of Personality Assessment, 87*, 121–138. doi:10.1207/s15327752jpa8702_02

Nichols, D. S., Padilla, J., & Gomez-Maqueo, E. L. (2000). Issues in the cross-cultural adaptation and use of the MMPI–2. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 247–292). Mahwah, NJ: Erlbaum.

Pascal, G. R., & Suttell, B. J. (1951). *The Bender-Gestalt test*. New York, NY: Grune & Stratton.

Peeters, M. A. G., Rutte, C. G., van Tuijl, H. F. J. M., & Reymen, I. M. M. J. (2006). The Big Five personality traits and individual satisfaction with the team. *Small Group Research, 37*, 187–211. doi:10.1177/1046496405285458

Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research, 7*, 156–176. doi:10.1177/074355489272003

Ponterotto, J. G. (2010). Multicultural personality: An evolving theory of optimal functioning in culturally heterogeneous societies. *The Counseling Psychologist, 38*, 714–758. doi:10.1177/0011000009359203

Porter, S. R., & Umbach, P. D. (2006). College major choice: An analysis of person-environment fit. *Research in Higher Education, 47*, 429–449. doi:10.1007/s11162-005-9002-3

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Journal of Personality Assessment, 22*, 203–212. doi:10.1037/a0018679

Rabin, A. I. (1986). *Projective techniques for adolescents and children.* New York, NY: Springer.

Rich, J. (2007). Psychological testing: Old specialty, new markets. *National Psychologist, 16*. Retrieved from http://nationalpsychologist.com/articles/art_v16n4_2.htm

Ridley, C. R., Li, L. C., & Hill, C. L. (1998). Multicultural assessment: Reexamination, reconceptualization, and practical application. *The Counseling Psychologist, 26*, 827–910. doi:10.1177/0011000098266001

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing.* New York, NY: Guilford Press.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180. doi:10.1016/j.hrmr.2006.03.004

Roysircar-Sodowsky, G., & Maestas, M. V. (2000). Acculturation, ethnic identity, and acculturative stress: Evidence and measurement. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 131–172). Mahwah, NJ: Erlbaum.

Ryan, J. J., & Lopez, S. J. (1996). Degree-granting institutions of diplomates and fellows in clinical neuropsychology. *The Clinical Neuropsychologist, 10*, 332.

Ryan, J. J., Lopez, S. J., & Lichtenberg, J. W. (1999). Neuropsychological training in APA-accredited counseling psychology programs. *The Counseling Psychologist, 27*, 435–442. doi:10.1177/0011000099273007

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178–200. doi:10.1037/h0023624

Schuerger, J. M. (2000). The sixteen personality factor questionnaire (16PF). In C. E. Watkins Jr. & V. L. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 73–110). Mahwah, NJ: Erlbaum.

Silvestri, T. J., & Richardson, T. Q. (2001). White racial identity statuses and NEO personality constructs: An exploratory analysis. *Journal of Counseling and Development, 79*, 68–76. doi:10.1002/j.1556-6676.2001.tb01945.x

Slaney, R. B., & MacKinnon-Slaney, F. (2000). Using vocational card sorts in career counseling. In C. E. Watkins Jr. & V. L. Campbell (Eds.), *Testing and assessment in counseling practice* (2nd ed., pp. 371–428). Mahwah, NJ: Erlbaum.

Smelson, D. A., Kordon, M. E., & Rudolph, B. (1997). Evaluating the diagnostic interview: Obstacles and future directions. *Journal of Clinical Psychology, 53*, 497–505. doi:10.1002/(SICI)1097-4679(199708)53:5<497::AID-JCLP12>3.0.CO;2-9

Snyder, D. K. (1997). *Marital Satisfaction Inventory—Revised.* Los Angeles, CA: Western Psychological Services.

Strong, E. K. (1935). Predictive value of the vocational interest test. *Journal of Educational Psychology, 26*, 331–349. doi:10.1037/h0062498

Sue, S. (1988). Psychotherapeutic services for ethnic minorities: Two decades of research findings. *American Psychologist, 43*, 301–308. doi:10.1037/0003-066X.43.4.301

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *MMPI–2 Restructured Clinical (RC) Scales: Development, validation, and interpretation.* Minneapolis: University of Minnesota Press.

Watkins, C. E., Jr. (1983). Counseling psychology versus clinical psychology: Further explorations on a theme or once more around the 'identity' maypole with gusto. *The Counseling Psychologist, 11*, 76–92. doi:10.1177/0011000083114012

Watkins, C. E., Jr., & Campbell, V. L. (1989). Personality assessment and counseling psychology. *Journal of Personality Assessment, 53*, 296–307.

Watkins, C. E., Jr., Campbell, V. L., Hollifield, J., & Duckworth, J. (1989). Projective techniques: Do they have a place in counseling psychology training? *The Counseling Psychologist, 17*, 511–513. doi:10.1177/0011000089173010

Watkins, C. E., Jr., Campbell, V. L., & Manus, M. (1990). Personality assessment training in counseling psychology programs. *Journal of Personality Assessment, 55*, 380–383. doi:10.1207/s15327752jpa5501&2_36

Watkins, C. E., Jr., Campbell, V. L., McGregor, P., & Godin, K. (1989). The MMPI: Does it have a place in counseling psychology training? *Journal of Personality Assessment, 53*, 413–417. doi:10.1207/s15327752jpa5302_17

Watkins, C. E., Jr., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*, 54–60. doi:10.1037/0735-7028.26.1.54

Williams, L., & Tappan, T. (1995). The utility of the Myers-Briggs perspective in couples counseling: A clinical framework. *American Journal of Family Therapy, 23*, 367–371. doi:10.1080/01926189508251367

Zimmerman, M. (2003). Integrating the assessment methods of researchers into routine clinical practice: The Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project. In M. B. First (Ed.), *Standardized evaluation in clinical practice* (pp. 29–74). Washington, DC: American Psychiatric Press.

CHAPTER 25

ASSESSMENTS OF PERCEIVED RACIAL STEREOTYPES, DISCRIMINATION, AND RACISM

Hyung Chol Yoo and Stephanie T. Pituc

The long-standing history of racism in the United States creates a significant gap in the quality of life between racial majority and minority groups. Racial disparities persist today in areas of law, education, employment, housing, media, health care, and health outcomes, to name a few (Feagin, 2000; Min, 2005; National Research Council, 2004; Takaki, 1993; U.S. Department of Health and Human Services, 2001). In the past 50 years, there has been exponential growth in research examining the experiences of, processes related to, and effect of racism in psychology. This chapter reviews measures of self-reported racial experiences, including perceived experiences of racial stereotypes, racial discrimination, and racism.

Given the proliferation of research studies examining racism, this chapter highlights theory-driven measures with strong psychometric support that has been published in peer-reviewed journals. There is a large extant literature examining the perpetrator's perspective in formation, maintenance, and behaviors of racism (for reviews, see Biernat & Crandall, 1999; Burkard, Medler, & Boticki, 2001). This chapter focuses instead on the experiences of the target—that is, the individual who is subject to racist treatment. First, this chapter provides a brief history on the definition of race and racism in the United States. Second, it elaborates on how race and racism has been studied in psychology, with particular attention to operational definitions and theoretical models of measurement. Third, it briefly reviews self-report measures with psychometric support, including an evaluation of each measure's strengths and weaknesses. Finally, the chapter concludes with thoughts on future directions in the field.

HISTORY AND CONSTRUCTION OF RACE AND RACISM IN THE UNITED STATES

Race is a sociopolitical construction based on perceived physical differences (e.g., skin color, facial features, hair type) and is often conflated or interchanged with ethnicity (i.e., group membership based on shared values, traditions, behaviors, and language; Betancourt & Lopez, 1993; Helms & Talleyrand, 1997). In the late 18th century, biologically based definitions of race served evolutionary theories' purposes to justify the racial inequality of minority groups (Gould, 1994; Lopez, 2006; Richards, 1997). However, science has generally debunked the notion of distinct biological and hierarchical differences between racial groups, finding more within-group than between-group variations in physical and biological characteristics (Betancourt & Lopez, 1993; Carter & Pieterse, 2005; Smedley & Smedley, 2005).

Nevertheless, invariably conceptualized within historical contexts, individuals are racialized, as race shapes group membership, meaning, experiences, and treatment of others (Helms, 1990; Helms & Cook, 1999; Kwan, 2005; Omi & Winant, 1994). Therefore, race and racism (i.e., a system of privilege and oppression based on racial hierarchy) are inextricably linked today and throughout history. Since the beginning of U.S. history, racism has often been intentionally blatant and violent toward racial

DOI: 10.1037/14048-025
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

minority groups, justified by false beliefs in a natural order of biological differences in which Whites were superior to people of color (Min, 2005; Takaki, 1993). These experiences include the genocide of Native Americans, slavery of African Americans, exclusions of Asian Americans, and colonization of Hispanics/Latinos and Pacific Islanders (Miller & Garran, 2008; Min, 2005; Takaki, 1993).

Definitions and classification of race and racial groups often change to maintain the status quo with regards to racial hierarchy. For instance, the U.S. Supreme Court used various definitions of race (e.g., biological vs. social) to deny citizenship to racial minorities (e.g., regarding cases *Takao Ozawa v. U.S.* and *Bhagat Singh Thind v. United States*; Lopez, 2006). Irish, Italian, and Jewish immigrants were once considered racial "others" and subjected to blatant forms of racism (Guglielmo & Salerno, 2003). Multiracials were classified and treated as non-Whites based on legislation and racist policies associated with the "one-drop rule" (Min, 2005; Root, 2000).

With the end of World War II and the beginning of the U.S. Civil Rights movement in the 1950s, there was another change in the tenor and practice of racism. It evolved from "old-fashioned" blatant, overt, and intentional expressions of White racial superiority into "modern" subtle, ambiguous, and unintentional reinforcement of the racial hierarchy (Devine, 1989; Dovidio & Gaertner, 1998; Plous, 2003; Sue, 2005). Racism today includes not only a conscious, active desire to hurt but also the omissions, inactions, and failure to help (Saucier, Miller, & Doucet, 2005). The model minority stereotype of Asian Americans as problemfree, industrious, and successful illustrates the complexity of modern-day racism. For Asians, this ostensibly positive image may lead to increased burden and pressure, negatively affecting identity and mental health (Yoo, Burrola, & Steger, 2010). For other racial groups, the stereotype perpetuates a White supremacist power structure in which non-Asian groups are denigrated for being unable to achieve the "success" of Asians in the United States (Wu, 2002).

Dictionary definitions and public connotations of racism center on "negative attitudes or behaviors in response to one racial group feeling superior to another" (e.g., "Racism," *Merriam-Webster's Collegiate Dictionary* [11th ed.], 2003; *Webster's Unabridged Dictionary*, 2005). This conceptualization, however, restricts the discourse on racism to individual, intentional, and negative acts of "meanness" (Sue, 2005; Tatum, 1997). Rather, racism is a system comprising individual, institutional, and cultural phenomena that change over time and history. It is a dynamic social and cultural construction, continually deconstructed and reconstructed to reinforce a power structure of privilege that benefits the racial majority over racial minority groups (Omi & Winant, 1994).

STUDY OF RACE AND RACISM IN PSYCHOLOGY

Psychologists have both promoted and fought against racism throughout the history of the discipline of psychology. The quality and characterization in psychological studies of race and racism can be divided into four broad overlapping periods: Scientific Racism (1860–1910), Race Psychology (1910–1940), Antiracism (1940–1970), and Race Relativism (1970–Present; Richards, 1997; Winston, 2004). Framed by the evolutionary thinking of underlying eugenicist scientific racism theories, many founders of U.S. psychology (including Sir Francis Galton, Herbert Spencer, and G. Stanley Hall) promoted natural racial differences, with the idea of White psychological faculties (e.g., intelligence, cognitive process, and development) as superior to the "savage," "barbaric," "less civilized," "primitive," and "adolescent" races. Studies during the Race Psychology period continued to assume White supremacy over non-Whites.

At the end of World War II and the beginning of the U.S. Civil Rights movement, the dialogue on race and racism in psychology substantively changed. During the Antiracism period, there was a more conscious recognition and discussion of racial bias. In the discipline of social psychology, in particular, critical and systematic investigation of racist attitudes and related dispositions began (e.g., Allport, 1954), with an emphasis on empirical studies and measurement focused on the perpetuator's perspective in understanding why, how, and when individuals were prejudiced (see Fiske, 1998, for a

review). This body of work led to increased understanding of individual differences (e.g., authoritarian personality), cognitive and motivational processes, and social group dynamics in the formation and maintenance of racism.

Despite a continually flourishing literature on prejudicial attitudes, the study of racial minorities' experiences of racism is a relatively recent phenomenon. The current period of Race Relativism brings new insights to the literature, empowering racial minorities to develop their own narratives in which they make meaning, struggle, and cope with racism (Sue et al., 2007). Reviews of empirical studies have begun to identify how and when racially stigmatized group members perceive encounters of racism (Crocker, Major, & Steele, 1998; Swim & Stangor, 1998) and the potential negative consequences on mental and physical health, adjustment, and substance use (Brondolo, Rieppi, Kelly, & Gerin, 2003; Gee, Ro, Shariff-Marco, & Chae, 2009; Paradies, 2006; Williams & Mohammed, 2009). Moreover, studies are beginning to identify specific ecological factors (e.g., person, family, community, and society) that can buffer or exacerbate the effects of racism (Brondolo, Brady, Pencille, Beatty, & Contrada, 2009; Gee et al., 2009).

In the past 50 years, there has been a rapid growth of research on race and racism. As demonstrated in Figure 25.1, the number of publications returned using PsycINFO increased from 115 of 87,934 total published peer-reviewed journals (0.13%) between 1960 and 1970 to 4,669 of 988,252 total published peer-reviewed journals (0.47%) between 2000 and 2010, using the key terms *racial stereotypes*, *racial discrimination*, *racial prejudice*, and *racism*. Unfortunately, the majority of empirical studies examining experiences and correlates of racism among racial minorities utilize racism measures without substantial validity and reliability evidence. Studies have often used single or few items to measure perceived racism (Karlsen & Nazroo, 2002; Noh & Kaspar, 2003) or modified other measures of discrimination without properly testing validity and reliability on interested samples (e.g., a measure based on African American experiences used in a study focusing on Asian Americans; Gee, Delva, & Takeuchi, 2007; Utsey, Chae, Brown, & Kelly, 2002). Consequently, the link between racism and outcome in these studies may be influenced by unknown measurement error, method variance, or poor construct validity.

In efforts to address some of these limitations, there has been significant advancement in developing more psychometrically rigorous instruments of perceived racism. In the first review of validated self-report measure of perceived racism, Utsey (1999) summarized six instruments that primarily focused on racialized experiences of African Americans. Today, there are considerably more instruments based on diverse theoretical models, racial group

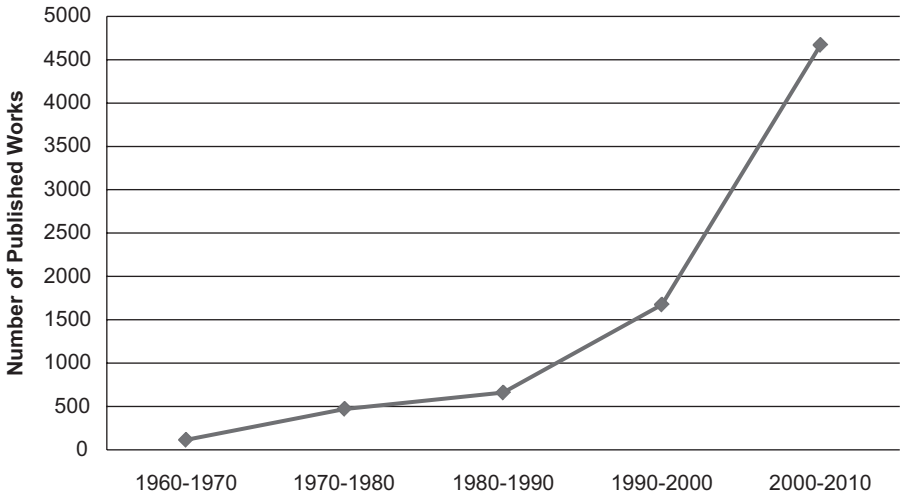


FIGURE 25.1. Number of peer-reviewed published works on racial stereotypes, racial discrimination, racial prejudice, and racism over the past 50 years.

populations, and quality of experiences measured (e.g., perceived, actual, frequency, and stressfulness). Our review builds on more recent reviews (Bastos, Celeste, Faerstein, & Barros, 2010; Kressin, Raymond, & Manze, 2008) in several important ways. First, our review focuses on instruments with evidence of construct validity published in peer-reviewed journals. Second, our review included measures within a broader framework of racism, including measuring racism as an acculturative stress, internalized racism, and stigma vulnerability. We believe that this point is particularly important to advance the field, as it more accurately reflects the complexity of racism. Third, we briefly summarize and critique each instrument to assist researchers in selecting measures most appropriate for their research studies.

The theories represented by the instruments reviewed include: (a) racism-related stress, (b) acculturative stress, (c) racial microaggression, (d) internalized racism, and (e) stigma vulnerability. Racism-related stress theory, the most popular framework in conceptualizing and measuring perceived racism, suggests that racism is a unique life stressor that can tax individuals' coping resources when perceived and, consequently, threaten mental and physical health (Harrell, 2000). Thus, these measures assess the frequency and appraisal of an event. Acculturative stress theory suggests that racism is a behavioral experience for many immigrants as a process of acculturation and adaptation to a new society (Sam & Berry, 2006). Thus, these measures assess the unique quality of immigrants not being able to fit in based on such characteristics as nationality or language. Racial microaggression theory suggests that contemporary experiences of racism are often subtle, brief, unconscious, chronic, and day-to-day indignities faced by people of color (Sue et al., 2007). Thus, these measures assess daily experiences of subtle racism. Internalized racism theory refers to the acceptance of distorted or inaccurate images, stereotypes, and societal beliefs by members of stigmatized racial groups (Jones, 1997). Many of these measures assess attitudes and the extent to which one agrees with false stereotypes and beliefs about one's group. Stigma vulnerability theory refers to the tendency to attribute negative interpersonal outcomes in ambiguous situations to racism by socially stigmatized groups (Crocker & Major, 1989). Thus, these measures assess the likelihood of attributing ambiguous interpersonal situations to racism. A summary of measures reviewed is provided in Table 25.1.

REVIEW OF SELF-REPORT MEASURES

Acculturative Stress Inventory for Children (ASIC)

The ASIC (Suarez-Morales, Dillon, & Szapocznik, 2007) is a 12-item, self-report measure using the items of the Societal, Attitudinal, Familial, and Environmental Acculturative Stress Scale for Children (Chavez, Moran, Reid, & Lopez, 1997), developed as a measure of acculturative stress among Hispanic children. Items are rated on a 6-point scale ranging from *doesn't apply* (0) to *bothers me a lot* (5). The ASIC has two subscales: Perceived Discrimination and Immigration-Related Experiences. Perceived Discrimination refers to perceived limitations as well as the feelings of being excluded or marginalized (eight items, e.g., the feeling of being overlooked by those who are supposed to provide help). Immigration-Related Experiences refer to feelings about living in a new country, being away from a country of origin, and speaking a new language.

The development of ASIC was guided by the acculturative stress theory, with the Perceived Discrimination subscale reflecting experiences of subtle racism, blatant racism, and exclusion, and marginalization. The items seem general enough to apply to other immigrant children, although scores from the instrument have been currently validated with only Hispanic children. Initial evidence of structural validity is good based on exploratory factor analysis (EFA) with clear rationale and choice of the two-factor structure; although as the test's authors admitted, confirmatory factor analysis (CFA) would greatly enhance the support of structural validity. As hypothesized, evidence of convergent and discriminant validity are supported with a moderate association (or lack thereof) with daily hassles, anxiety, social desirability, and racial group differences. Also, given the relatively large effect size between their measure and a daily hassles measure ($r = .57$), it

TABLE 25.1

Summary of Racism Measures

Scale name	Theory	No. of items	Subscales	Population	Strengths	Weaknesses
Acculturative Stress Inventory for Children (ASIC; Suarez-Morales et al., 2007)	Acculturative stress	12	Perceived Discrimination and Immigration-Related Experiences	Hispanic children	One of few measures examining experiences of immigrant children. Good evidence of convergent and discriminant validity. Good reliability.	Items general enough to apply to other immigrant children, but only validated with Hispanic children.
Adolescent Discrimination Distress Index (ADDI; Fisher et al., 2000)	Racism-related stress	15	Institutional Discrimination Distress, Educational Discrimination Distress, and Peer Discrimination Distress	African American, Asian American, Hispanic, White American adolescents	Focuses on uniquely racialized experiences of minority adolescents across racial groups.	Evidence of structural validity not clear Incremental validity evidence not clear.
American-International Relations Scale (AIRS; Sodowsky & Plake, 1991b)	Acculturative stress	34	Perceived Prejudice, Social Customs, and Language Use	Asian, African, and South American international college students, faculty, and staff	One of few measures for international students. One of few measures of discrimination that can be used with different international racial groups.	Evidence of structural validity not clear.
Asian American Racism-related Stress Inventory (AARRSI; Liang et al., 2004)	Racism-related stress	29	Socio-Historical Racism, General Racism, and Perpetual Foreigner Racism	Asian American college students	One of few measures tapping into unique racism-related stress of Asian Americans. Good evidence of structural validity.	Scale format may conflate frequency and stressfulness of an event. No evidence of incremental validity.
Color-Blind Racial Attitude Scale (CoBRAS; Neville et al., 2000)	Internalized racism	20	Unawareness to Racial Privilege, Institutional Discrimination, and Blatant Racial Issues	White, African American, American Indian, Asian American, Latino college students, community adults	One of few measures assessing denial in existence of racism. Good evidence of construct validity.	Subscales of CoBRAS (in particular, Blatant Racial Issues) may not be stable over time.
Everyday Discrimination Scale (EDS; Williams et al., 1997)	Racial microaggression	9	Unidimensional	Black and White community adults	Popular measure in epidemiological studies. Only measure reviewed that captured subtle, brief, and day-to-day indignities faced by individuals of color.	Efforts to generalize measure across racial groups may be limited by some of its items specific to particular racial groups' racialized experiences.

(*Continued*)

TABLE 25.1 (*Continued*)

Summary of Racism Measures

Scale name	Theory	No. of items	Subscales	Population	Strengths	Weaknesses
Experiences of Discrimination Measure (EOD; Krieger, Smith, Naishadham, Hartman, & Barbeau, 2005)	Racism-related stress	9	Unidimensional	Black and Latino community adults	Popular measure in epidemiological studies. Only measure that assesses actual count of type of discrimination experienced rather than perceived experiences. Good evidence of structural validity.	Items are general enough to use with diverse racial groups but currently limited to validation with only Black and Latino adults.
General Ethnic Discrimination Scale (GEDS; Landrine, Klonoff, Corral, Fernandez, & Roesch, 2006)	Racism-related stress	18	Recent (Past Year), Lifetime (Entire Life), and Appraised (Stressful)	Black, Latino, Asian, and White college and community adults	Valid across diverse racial groups. Generally, short in assessing both frequency and stressfulness in range of individual experiences of discrimination. Good evidence of structural validity and reliability.	Items mix both experiences and emotional and behavioral responses to discrimination. Initial study primarily conducted on a female population.
Historical Loss Scale (HLS; Whitbeck et al., 2004)	Racism-related stress	12	Unidimensional	American Indian adults	Short. Only instrument measuring historical racism. Only instrument for Native Americans.	Generalizability of measure across the many Native American tribes is not clear.
Index of Race-Related Stress (IRRS; Utsey & Ponterotto, 1996)	Racism-related stress	46	Individual Racism, Cultural Racism, Collective Racism, and Institutional Racism	Children aged 10–12 years	One of few measures that assess perceived discrimination at group levels.	Lengthy. Scale format may conflate frequency and stressfulness of an event. Incremental validity evidence not clear.
Index of Race-Related Stress-Brief (IRRS-Brief; Utsey, 1999)	Racism-related stress	22	Individual Racism, Cultural Racism, and Institutional Racism	African American college and community adults	Shorter than the original measure of IRRS.	Scale may conflate frequency and stressfulness of an event. Incremental validity evidence not clear.
Index of Race-Related Stress-Adolescents (IRRS-Adolescents; Seaton, 2003, 2006)	Racism-related stress	32	Individual Racism, Cultural Racism, and Collective/Institutional Racism	Black adolescents	Able to measure multilevels of racism-related stress with Black adolescent populations.	Scale format may conflate frequency and stressfulness of an event. Incremental validity evidence not clear.

Internalization of the Model Minority Myth Measure (IM-4; Yoo, Steger, & Castro, 2010)	Internalized racism	15	Model Minority Myth of Achievement Orientation and Model Minority Myth of Unrestricted Mobility	Asian American college students	One of few measures assessing uniquely racialized experiences of Asian Americans. Only measure reviewed assessing positive dimension of racism.	Preliminary evidence does not suggest clear or strong evidence of convergent and discriminant validity. Generalizability of the instrument currently limited to academically successful Asian American college students.
Minority Status Stress Scale (MSS; Smedley et al., 1993)	Racism-related stress	33	Social Climate Stress, Interracial Stresses, Racism and Discrimination, Within-Group Stress, Achievement Stresses	African-American, Chicano, Latino, and Filipino college students	Only measure assessing racism-related stress associated with within-group discrimination.	Evidence of structural validity of the measure across different racial minority groups is not clear.
Perceived Ethnic Discrimination Questionnaire (PEDQ-Original; Contrada et al., 2001)	Racism-related stress	17	Disvaluation, Threat/ Aggression, Verbal Rejection, and Avoidance	White (mostly), African American, Hispanic, and Asian/Pacific Islander college students	One of few measures assessing the perceived frequency in multiple dimensions of interpersonal, negative, racial discrimination. Good evidence of convergent and discriminant validity.	Possibly limited to college sample. Small sample of racial minority group. Disvaluation subscale may not reflect experiences of Asian Americans.
Perceived Ethnic Discrimination Questionnaire-Community Version (PEDQ-CV; Brondolo et al., 2005)	Racism-related stress	70	Exclusion/Rejection, Stigmatization/ Disvaluation, School/ Work Discrimination, Threat/Aggression, Media Discrimination, Discrimination Against Family Member, Discrimination in Different Settings, Past Week Discrimination	Black (mostly), Latino (mostly), White, Asian, Native American, and Multiracial community adults	One of few measures assessing the perceived frequency in multiple dimensions of interpersonal, negative, racial discrimination. Increased generalizability in validation on Black and Latino community adults.	Lengthy. Several scales in measure not validated. Several items across subscales do not reflect discrimination experiences of Asian Americans.
Perceived Ethnic Discrimination Questionnaire-Brief Version (Brief PEDQ-CV; Brondolo et al., 2005)	Racism-related stress	17	Exclusion/Rejection, Stigmatization/ Disvaluation, School/ Work Discrimination, Threat/Aggression	Black (mostly), Latino (mostly), White, Asian, Native American, and Multiracial community adults and college students	Short. One of few measures assessing frequency in multiple dimensions of interpersonal, negative, racial discrimination. Increased generalizability in validation on Black and Latino community adults.	Additional structural validity and factor structure need to be evaluated. Several items across subscales may not reflect experiences of Asian Americans.

(*Continued*)

TABLE 25.1 (*Continued*)

Summary of Racism Measures

Scale name	Theory	No. of items	Subscales	Population	Strengths	Weaknesses
Perceived Racism Scale (PRS; McNeilly et al., 1996)	Racism-related stress	51	Frequency of Exposure Subscales include: Employment, Academic, Public Realm, and Racist Statements; Emotional Response Subscales include: Anger/ Frustration, Depressed Affect, and Feeling Strengthened; Behavioral Coping Response Subscales include: Working Harder/Trying to Change Things, Avoiding/ Ignoring, Praying, Forgetting It, Getting Violent, and Speaking Up	African American college students and community adults	Comprehensive measure of frequency, emotional response, and behavioral response of uniquely racialized experiences of discrimination faced by African Americans.	Lengthy. Structural validity evidence is compromised given sample size and actual number of items.
Perceptions of Racism in Children and Youth (PRaCY; Pachter et al., 2010)	Racism-related stress	10	Unidimensional	Latino, African American, West Indian/Caribbean, and multiracial youths	Good evidence of structural validity. One of few measures intended for children and adolescents from diverse racial minority groups.	Some items may not be relevant across all racial minority groups; consequently, further cross-validation is necessary.
Prejudice Perception Assessment Scale (PPAS; Gilbert, 1998)	Stigma vulnerability	5	Unidimensional	African American college students	Short. One of few measures assessing the extent to which individuals are likely to attribute ambiguous cross-racial interactions to racial prejudice.	Limited to assessment of stigma consciousness for African Americans in predominantly White universities.
Race-Related Stressor Scale (RRSS; Loo et al., 2001)	Racism-related stress	33	Racial Prejudice and Stigmatization, Bicultural Identification and Conflict, and a Racist Environment	Asian American Vietnam veterans	Only measure assessing stressors faced by Asian American Vietnam war veterans. Good evidence of predictive validity on mental health.	Structural validity evidence and differentiation of subscales is not clear.

Scale of Ethnic Experience (SEE; Malcarne, Chavira, Fernandez, & Liu, 2006)	Acculturative stress	32	Ethnic Identity, Perceived Discrimination, Mainstream Comfort, and Social Affiliation	African American, White, Filipino Americans, and Mexican American college students	Discrimination items are broad enough to be used across ethnic groups.	Test–retest reliability for Perceived Disc subscale was low for African Americans. Potentially conflates individual and group discrimination.
Schedule of Racist Events (SRE; Landrine & Klonoff, 1996)	Racism-related stress	18	Recent (Past Year), Lifetime (Entire Life), and Appraised (Stressful)	African American college students and community adults	Assesses unique racial discrimination experiences and stressfulness of African Americans. Good evidence of validity and reliability.	Items mix both experiences and emotional and behavioral responses to discrimination.
Subtle and Blatant Racism Scale for Asian Americans (SABR-A^2; Yoo, Steger, & Lee, 2010)	Racism-related stress	8	Subtle Racism and Blatant Racism	Asian American college students	Assesses unique subtle and blatant racialized experiences of Asian Americans. Short, easy, and quick to administer.	Does not capture the diverse spectrum of unique racism experiences of Asian Americans.

would have been more convincing to demonstrate evidence of incremental validity by examining the relationship between acculturative stress and anxiety, controlling for daily hassles. Internal consistency reliability and 2-week test–retest reliability are both adequate (above .70).

Adolescent Discrimination Distress Index (ADDI)

The ADDI (Fisher, Wallace, & Fenton, 2000) is a 15-item self-report measure of adolescent distress in response to perceived instances of racially motivated discrimination in institutional (e.g., stores, restaurants), educational (e.g., teacher evaluations), and peer contexts.

After each statement, students are asked to indicate whether they experienced the type of discrimination because of their race or ethnicity, and if they have, to rate how much it upset them, on a 5-point scale ranging from *not at all* to *extremely*. The ADDI has three subscales: Institutional Discrimination Distress, Educational Discrimination Distress, and Peer Discrimination Distress. Institutional Discrimination Distress refers to distress associated with institutional discrimination (six items; e.g., police harassment). Educational Discrimination Distress refers to distress associated with educational discrimination (four items; e.g., being discouraged to enroll in advanced coursework). Peer Discrimination Distress refers to distress associated with peer discrimination (five items; e.g., exclusion from group activities).

The development of the ADDI was guided by the racism-related stress theory and is a popular measure that taps into different levels of racism (including individual and institutional) and can be used with different racial groups (i.e., African American, East Asian, South Asian, Hispanic, and White American adolescents). It is one of the few measures for which evidence of validity has been collected for the use of the scales with minority adolescents. Its authors assessed racial group variations in estimations of validity and reliability. Although they conducted a hypothesized three-factor principal-components analysis (PCA) in assessment of structural validity, it is not clear whether this model is the best fit. Several items that were retained loaded highly across more than one factor but the possibility of an alternative model fit was not explored. Additionally, a CFA to examine evidence of structural validity was not conducted. The evidence of convergent and discriminant validity is acceptable; as expected correlations between the ADDI subscale scores and racial bias preparation, self-esteem, and developmental variation are small. However, evidence of incremental validity of the instrument was not assessed, which seems particularly relevant for racism-related stress measures. For instance, how do we know whether the small correlation between the ADDI subscales and self-esteem is attributed to specific discrimination distress or to general distress? Internal consistency reliability and test–retest reliability are modest, with two of three subscales' internal consistency values at .60.

American-International Relations Scale (AIRS)

The AIRS (Sodowsky & Plake, 1991b) is a 34-item self-report measure of adjustment between the dominant culture and original culture of Asian, African, and South American international college students, faculty, and staff. It is rated on a 6-point scale ranging from *strong disagreement* (1) to *strong agreement* (6). AIRS has three subscales: Perceived Prejudice, Acculturation, and Language Usage. Perceived Prejudice refers to the degree of acceptance by Americans of international people (20 items; e.g., feeling resentful about a lack of recognition). Acculturation (11 items) refers to degree of acceptance of Americans and American culture. Language Usage (three items) refers to the language usage, proficiency, and preference of international people.

The development of AIRS was guided by the acculturative stress theory. It is one of few measures that assesses acculturative stress and adjustment of international students from different racial backgrounds and discusses specific racialized discriminatory experiences of international students (e.g., being treated like a foreigner despite acculturation). Evidence of structural validity of the instrument is questionable. There is some evidence of validity of the instrument in mental health contexts (Atri, Sharma, & Cottrell, 2006–2007) and evidence of criterion-related validity with expected differences

across international groups, permanent versus nonpermanent U.S. resident status, and length of residence in the United States (Sodowsky & Plake, 1992). Internal consistency reliability values of the instrument and its subscales are good (.79–.89). The effort has been made to generalize the measure with Asian Americans and Hispanics, renamed the Majority-Minority Relations Survey (Sodowsky & Plake, 1991a); however, the three-factor model was a poor fit based on CFA and a goodness-of-fit index reported at .73.

Asian American Racism-Related Stress Inventory (AARRSI)

The AARRSI (Liang, Li, & Kim, 2004) is a 29-item self-report measure of racism-related stress among Asian American college students. It is rated on a 5-point scale ranging from *this has never happened to me or someone I know* (1) to *this event happened and I was extremely upset* (5). AARSI has three subscales: Socio-Historical Racism, General Racism, and Perpetual Foreigner Racism. Socio-Historical Racism refers to collective experiences, transgenerational transmission, vicarious racism experiences, and chronic contextual stress of racism-related stress for Asian Americans (14 items; e.g., noticing the omission of Asian Americans from U.S. history books). General Racism refers to daily stressors and racism life event types (eight items; e.g., people assuming that you are good at math). Perpetual Foreigner Racism refers to the stereotype of Asian Americans as perpetual foreigners (seven items; e.g., being spoken to in an unnaturally slow manner).

The development of the AARRSI was guided by racism-related stress theory. It is one of few measures that examine stress related to the unique experiences of Asian Americans, including assumptions of Asian Americans as foreign, unable to speak English, and being overachievers. Its authors developed the measure using both EFA and CFA on two independent samples to find good evidence of structural validity of a three-factor model. Evidence of concurrent validity is demonstrated with hypothesized significant positive correlations between AARRSI total and subscale scores and other measures of racism, despite finding no relationship with psychological adjustment. However, AARRSI total scores correlated with career problems, self-esteem problems, and interpersonal problems in a separate study (Liang & Fassinger, 2008). Evidence of discriminant validity is supported with an expected nonsignificant relationship between scores on AARRSI and Asian values. The scale format potentially confounds the frequency of events with the perceived stressfulness of the events. Given the measurement of stressfulness, it would have been a plus to demonstrate evidence of incremental validity of AARSI on psychological adjustment beyond general stress. Internal consistency reliability of the instrument and its subscales ranged from adequate to very good (.75–.95). Two-week test–retest reliability coefficients are adequate to good (.73–.87).

Color-Blind Racial Attitude Scale (CoBRAS)

The CoBRAS (Neville, Lilly, Duran, Lee, & Browne, 2000) is a 20-item self-report measure in the denial of the existence of racism among White, African Americans, American Indian, Asian American, and Latino college students and community adults. It is rated on a 6-point scale ranging from *strongly disagree* (1) to *strongly agree* (6). CoBRAS has three subscales: Unawareness of Racial Privilege, Institutional Discrimination, and Blatant Racial Issues. Unawareness to Racial Privilege refers to blindness to the existence of White privilege in the United States (seven items; e.g., advantages that White people in the United States have as a result of skin color). Institutional Discrimination (seven items) refers to limited awareness of the implications of institutional forms of racial discrimination and exclusion. Blatant Racial Issues (six items) refers to a participant's unawareness of general, pervasive racial discrimination.

Informed by internalized racism theory, the CoBRAS is one of the few measures that assesses marginalized racial group members' belief that racism does not exist. The three-factor structure and item development provide good evidence of structural validity from both PCA and CFA on independent samples. Expected correlations with belief in a just world, racial prejudice toward minorities, and social desirability provide evidence of concurrent and discriminant validity. Moreover, evidence of

criterion-related validity is established with expected differences between gender and racial groups. Internal consistency reliability of the instrument and its subscales range from adequate to very good (.70–.91). Two-week test–retest reliability estimates for the Racial Privilege and Institutional Discrimination subscales are good (.80); however, 2-week test–retest reliability estimates of .34 for the Blatant Racial Issues subscale and .68 for the total score suggest possible temporal instability.

Everyday Discrimination Scale (EDS)

The EDS (Williams, Yu, Jackson, & Anderson, 1997) is a nine-item self-report measure of chronic, routine, and day-to-day experiences of general, unfair treatment among primarily African American community adults. Participants assess the frequency of occurrence of each item in everyday life, rating each item on a 4-point scale ranging from *never* to *often*. Participants then indicate the main reason for their experiences (e.g., ethnicity, gender). The EDS was originally used as a unidimensional scale, including experiences of being treated with less courtesy, poorer services in restaurants or stores, acting as if one is not smart, and being afraid.

The EDS was originally developed for the Detroit Area Study (Williams et al., 1997), which examined racial differences between Blacks and Whites in association with socioeconomic statuses, discrimination, and physical/mental health. The questions were based on previous qualitative studies of subtle, day-to-day microaggressions (Essed, 1991; Feagin, 1991). Its authors did not discuss measurement development, report psychometric properties (excluding Cronbach's alpha at .88), or instructions on how to calculate scores for those interested in a specific type of discrimination (e.g., racial). However, this instrument is popular in epidemiological studies with frequent use in large community and national studies with primarily African American and Latino adults (e.g., Guyll, Matthews, & Bromberger, 2001; Kessler, Mickelson, & Williams, 1999; Krieger, Smith, Naishadham, Hartman, & Barbeau, 2005; Williams et al., 1997) and Asian American adults (Gee, Spencer, Chen, & Takeuchi, 2007). Despite its popularity, psychometric evidence for the EDS is limited. Thus far, only a few papers briefly describe evidence of structural validity of the instrument supporting a one-factor solution (Clark, Coleman, & Novak, 2004; Gee, Spencer, Chen, & Takeuchi, 2007; Krieger et al., 2005) and a two-factor solution (Guyll, Matthews, & Bromberger, 2001). Scores from the EDS generally support evidence of predictive validity of mental and physical health outcomes, and internal consistency reliability is adequate (above .70). The EDS is one of the few measures that is based on racial microaggression theory (Sue et al., 2007). However, use of the EDS with some racial minority groups should be cautioned, as several of its items may not be applicable across racial groups (e.g., stereotypes of criminal behavior or inferior intellect).

Experiences of Discrimination (EOD)

The EOD (Krieger et al., 2005) is a nine-item self-report measure of racial discrimination in different settings (such as school, work, housing, medical care, and store or restaurant) among African American and Latino community adults. After each statement, participants are asked to indicate whether they experienced the type of discrimination and to rate the frequency on a 4-point scale ranging from *never* (4) to *four or more times* (1).

The development of EOD was guided by the racism-related stress theory. The EOD is another popular measure in epidemiological studies with frequent use in large community and national studies focusing on Black and Latino adult samples (e.g., Krieger & Sidney, 1996; Stancil, Hertz-Picciotto, Schramm, & Watt-Morse, 2000; Stuber, Galea, Ahern, Blaney, & Fuller, 2003). The test's authors emphasize the value of this instrument is that it captures discrimination in different settings (e.g., stores, school, work, etc.) and actual experiences rather than perceived experiences. The EOD scale scores demonstrate acceptable structural validity evidence of the one-factor model based on CFA results. Evidence of concurrent and discriminant validity also is demonstrated with expected correlations with other measures of discrimination, psychological stress, and social desirability. Internal consistency reliability (above .74) and test–retest reliability estimates are adequate (.70).

General Ethnic Discrimination Scale (GEDS)

The GEDS (Landrine, Klonoff, Corral, Fernandez, & Roesch, 2006) is an 18-item self-report measure of perceived ethnic discrimination. After each statement, participants are asked to indicate how often they have experienced the type of discrimination in their lifetime and within the past year, each on a 6-point scale ranging from *never* (1) to *almost all the time* (6). In addition, they are asked the level of stressfulness for each event on a 6-point scale ranging from *not at all stressful* to *extremely stressful*. The three ratings are treated separately as three different subscales, Recent (Past Year), Lifetime (Entire Life), and Appraised (Stressful).

The development of the GEDS was guided by the racism-related stress theory. It is almost identical in number of items and scale format to the popular Schedule of Racist Events for African Americans (Landrine & Klonoff, 1996) but reworded to generalize experiences of discrimination across diverse racial groups. The GEDS captures different sources of discrimination (e.g., teachers, neighbors), major events (e.g., racial slurs), daily hassles (e.g., one's motivations being misinterpreted), and behavioral and emotional reactions to discrimination (e.g., feeling anger or holding back from confrontation). It was specifically constructed for use across racial groups, and initial reports of evidence of structural validity (based on CFA, structural equation modeling, and multiple group analysis results) is strong with high factor loadings and good fit indexes with African American, Latin American, White American, and Asian American college and community adult populations. Convergent validity is demonstrated with expected correlations between GEDS subscale scores and scores of other measures of racial discrimination, psychiatric symptoms, and substance use (i.e., cigarette smoking). Internal consistency estimates for the GEDS subscales are very good (above .90) for each racial group. The GEDS shows promise as a quick, valid, and reliable instrument for scholars wishing to study discrimination across racial groups tapping into different qualities of individual level racism. Although analysis of the instrument's scores provides evidence for structural validity, the items represent a mix of exposure to experiences of discrimination and responses to discrimination. This conflation potentially compromises the argument for strong construct validity. Moreover, an overwhelming number of participants in this study were women (72%), and its generalizability across gender should be further examined in future studies.

Historical Loss Scale (HLS)

The HLS (Whitbeck, Adams, Hoyt, & Chen, 2004) is a unidimensional, 12-item self-report measure of perceived historical racism and the frequency with which reminders occur in loss of language, culture, land, and broken treaty promises faced by American Indians. Items are rated on a 6-point scale ranging from *never* (6) to *several times a day* (1).

The development of HLS was guided by the racism-related stress theory as the reminder of the loss associated with the historical legacy of racism-related trauma, stress, and unresolved grief experienced by Native Americans today (Belcourt-Dittloff & Stewart, 2000). As Harrell (2000) pointed out in her multidimensional conceptualization of racism-related stress, perceptions of historical racism or racism experienced by transgenerational transmission can be quite stressful reliving or being reminded of the historical trauma faced by one's minority group. Such experiences can include the slavery of African people, the internment of Japanese Americans during World War II, the removal of American Indians from their tribal lands, and refugee experiences (Root, 1993). The HLS is the only published instrument that measures the dimension of historical racism. Moreover, it is the only instrument reviewed assessing the uniquely racialized discrimination experience of Native Americans. In fact, most measures of general racism intended for diverse racial groups rarely include a substantive sample size or cross-validation with Native Americans. Evidence of structural validity of the instrument is assessed using EFA, with results suggesting a clear one-factor solution with high factor loadings (all above .62) of all items. Evidence of convergent validity is evaluated with expected positive correlations with anxiety/depression and anger/avoidance. Internal consistency reliability is very good (.92).

Index of Race-Related Stress (IRSS)

The original IRSS (Utsey & Ponterotto, 1996) is a 46-item self-report measure of stress experienced due to generally blatant encounters with racism faced by African American college students and community residents. Items are rated on a 5-point scale ranging from *this has never happened to me* (0) to *event happened and I was extremely upset* (4). IRSS has four subscales: Individual Racism, Cultural Racism, Collective Racism, and Institutional Racism. Individual Racism refers to the experience of racism on a personal level (11 items; e.g., being treated as if you are unintelligent). Cultural Racism refers to the cultural practices of one group being lauded as superior to those of another (16 items; e.g., noticing that certain acts such as crimes are viewed differently when done by Blacks vs. Whites). Collective Racism refers to when organized or semiorganized racial groups restrict the rights of other racial groups (eight items; e.g., trouble getting a cab). Institutional Racism refers to experiences as a result of racism being embedded in the policies of a given institution (11 items; e.g., being refused housing).

The IRRS's development was theoretically grounded in racism-related stress theory and psychometrically rigorous in assessment of construct validity. Structural validity of the instrument was assessed using both PCA and CFA on independent samples with clear rationale of item and factor retention. The initial CFA on the four-factor oblique model suggested a poor fit based on multiple fit indexes (e.g., goodness of fit index = .78), although reanalysis of the data using aggregate variables of existing items reduced random error associated with the factor and increased fit indexes (e.g., goodness of fit index = .90). The IRRS total and subscale scores generally demonstrate evidence of concurrent validity with other measures of racism and general stress, except the Collective Racism subscale. Evidence of criterion-related validity shows expected mean differences between Black and Whites. The IRRS is a comprehensive, multidimensional measure of racism-related stress. It is one of the few measures assessing perceived group discrimination, (i.e., institutional and cultural racism). However, the original IRRS is lengthy. Also, its scale format combines frequency and stressfulness; this ratings approach may confound frequency of events with the perceived stressfulness of the event. Internal consistency reliability of the instrument and its subscales range from adequate to good (.74–.89). Test–retest reliability estimates range from modest to adequate (.58–.79) with several subscales suggesting temporal instability.

In addressing some of the limitations of the original IRRS, Utsey (1999) developed the IRRS-Brief with 22 items that supported the original three subscales: Individual Racism, Cultural Racism, and Institutional Racism. The Collective Racism subscale was removed as a consequence of more stringent item selection criteria. Assessment of construct validity and reliability of the IRRS-Brief was comparable with the original IRRS. More recently, Seaton (2003) provided evidence of validity for the three-factor structure of the IRRS with a sample of African American adolescents.

Internalization of the Model Minority Myth Measure (IM-4)

The IM-4 (Yoo, Steger, & Lee, 2010) is a 15-item self-report measure of the extent to which individuals believe Asian Americans are more successful than other racial minority groups based on values emphasizing achievement and hard work and belief in unrestricted mobility toward progress. Items are rated on a 7-point scale ranging from *strongly disagree* (1) to *strongly agree* (7). IM-4 has two subscales: Model Minority Myth of Achievement Orientation and Model Minority of Unrestricted Mobility. Model Minority Myth of Achievement Orientation refers to the assumption of Asian Americans' greater success than other racial minority groups is associated with stronger work ethic, perseverance, and drive to succeed (10 items; e.g., believing that Asian Americans have higher grade point averages because they work harder). Model Minority of Unrestricted Mobility refers to beliefs that Asian Americans' greater success than other racial minority groups is associated with a stronger belief in meritocracy and lack of perceived racism or barriers at school/work (five items; e.g., believing that Asian Americans encounter less racial discrimination).

The development of the IM-4 was guided by internalized racism theory and emphasizes

endorsement of a uniquely racialized, positive, but distorted stereotype of Asian Americans. Its authors contend internalization of positive stereotypes and racism can lead to adverse psychological effects, similar to more common negative stereotypes and racism. It is one of the few measures of internalized racism and the only measure reviewed emphasizing a positive dimension of racism. Evidence of the structural validity of the two-factor model is supported by both EFA and CFA using independent samples. Yoo et al. (2010) also used a separate parallel analysis (in addition to interpretability, screeplot, and eigenvalues) to determine the number of factors. The evidence of convergent and discriminant validity are not as strong. In partial support of convergent validity, there are some significant relations between IM-4 subscale scores and ethnic identity components and situational well-being. In partial support of discriminant validity, there are small positive or nonsignificant relations between IM-4 subscale scores and Asian American values. In partial support of incremental validity, IM-4 subscales relate to psychological distress symptoms, even after controlling for Asian American values and ethnic identity components. The IM-4's generalizability may be limited, as evidence of validity is available only for academically successful college students. Internal consistency reliability values of the IM-4's subscales range from adequate to very good (.75–.91) and 2-week test–retest reliability coefficients are adequate (.70–.72).

Minority Status Stress Scale (MSS)

The MSS (Smedley, Myers, & Harrell, 1993) is a 33-item self-report measure of perceived stress attributed to being a racial minority in a predominantly White college and university setting. Scores from this measure were initially validated on a sample of African-American, Chicano, Latino, and Filipino college students. Items are rated on a 6-point scale ranging from *does not apply* (0) to *extremely stressful* (5). The MSS has five subscales: Social Climate Stresses, Interracial Stresses, Racism and Discrimination Stresses, Within-Group Stresses, and Achievement Stresses. Social Climate refers to stressors associated with campus climate (11 items; e.g., feeling that the university does not have concern for the needs of one's racial group). Interracial Stresses refer to stressors associated with problems managing relationships both within and outside of one's racial and ethnic group (seven items; e.g., having negative relationships with different ethnic groups at the university). Racism and Discrimination Stresses refer to stressors associated with racial discrimination (five items; e.g., unfair treatment due to one's race). Within-Group Stresses refer to stressors associated with discrimination and pressure to conform from within-group racial members (four items; e.g., showing loyalty to one's own race). Achievement Stresses refer to stressors associated with one's ability to succeed in college (six items; e.g., family expectations for academic success).

The development of the MSS was guided by racism-related stress theory. It is a popular measure used widely in research on minority college students' stress and its relationships with various psychological and academic outcomes. There is initial evidence of structural validity of MSS and its five-factor model based on PCA results. MSS subscale scores (in particular, the Achievement Stresses subscale) significantly contribute variance of psychological distress and academic performance (although not general well-being) beyond effects of race, gender, socioeconomic status, prior levels of academic preparation, and generic student stresses. Internal consistency reliability estimates range from adequate to very good (.76–.93).

Subscales' scores of the MSS uniquely tap into experiences of racial minorities in a college and university setting. A particular new dimension of racism offered by the MSS is the stress associated with within-group racism (i.e., perceived discrimination by other members of the same race). Too often, the discourse of racism is polarized as a Black and White issue, with many perceived racism measures assuming that the perpetuator is White. The Within-Group Stresses subscale of the MSS may provide insight into differential psychological effects of between- versus within-group racism-related stress. Although its authors made an effort to provide evidence of validity for the instrument's scores across racial minority groups, sample sizes of groups were too small to conduct any significant between-group analyses. Moreover, some items may not generalize

across racial minority groups. For instance, the item related to expectation of poor academic performance may not apply to racialized experiences of Asian Americans, for whom expectations run toward the other direction.

Perceived Ethnic Discrimination Questionnaire (PEDQ)

The PEDQ-Original (Contrada et al., 2001) is a 17-item self-report measure of interpersonal, individual level of racial discrimination among White, African American, Asian American, and Hispanic/Latino college students. Specifically, items reflect a wide range of racial discrimination, including verbal rejection, avoidance, exclusion, denial of equal treatment, disvaluing action, threat of aggression, and aggression. In response to each of these items, respondents are instructed to use a 7-point scale ranging from *never* (1) to *very often* (7) to indicate prevalence over the past 3 months. The PEDQ-Original has four subscales: Disvaluation, Threat/Aggression, Verbal Rejection, and Avoidance. Disvaluation refers to experiences of being treated as inferior based on race (six items; e.g., assumptions that you are dangerous). Threat/Aggression (five items) refers to experiences of being physically assaulted or threatened based on race. Verbal Rejection (three items) refers to experiences of verbal harassment based on race. Avoidance (three items) refers to race-based experiences of avoidance from others.

The development of the PEDQ-Original was guided by racism-related stress theory. It was constructed with the specific intention for use across diverse racial and ethnic groups. Consequently, items of interpersonal, negative racial discrimination were written broadly to capture racialized experiences across groups. However, the initial development sample of the PEDQ-Original was primarily White ($n = 208$), followed by Asian/Pacific Islander ($n = 60$), African American/Black ($n = 34$) and Hispanic/Latino ($n = 31$) college students. Moreover, items from the Disvaluation subscale seem to represent experiences counter to the model minority image of Asian Americans.

The initial report of evidence of structural validity is demonstrated using EFA. In support of evidence of criterion-related validity, mean score differences of PEDQ-Original subscale scores are found between Whites and people of color. In partial support of discriminant validity, PEDQ-Original subscale scores correlate very weakly with an ethnic identity measure. In support of convergent and partial support of incremental validity, total score correlate with mental health (i.e., negative mood, depression, and life satisfaction) and physical health (i.e., physical symptoms and health care visits). However, only its effect on depression remains after controlling for related variables (i.e., own-group conformity pressure, stereotype confirmation concern, generic stress, parents education, ethnic identity, and personal self-esteem). Internal consistency reliability estimates range from adequate to good (.74–.89).

On the basis of some limitations discussed earlier, Brondolo et al. (2005) developed a modified version of the PEDQ-Original that emphasized the interpersonal, individual levels of racial discrimination experienced by primarily Black and Latino community adults and college students (i.e., PEDQ-Community Version [PEDQ-CV] and Brief PEDQ-CV). Although the original measure inquired about a variety of everyday experiences broadly relevant to members of minority groups in general (Contrada et al., 2001), the community version emphasized the life experiences of community-dwelling adults. Moreover, Brondolo et al. added additional items to the full PEDQ-CV (four additional scales, including Discrimination in the Media, Discrimination against Family Members, Discrimination in Different Settings, and Past Week Discrimination) totaling 70 items. However, because the test's authors did not conduct a rigorous empirical investigation of these other scales, this review focuses on the Brief PEDQ-CV, which is the most comparable with the PEDQ-Original and most popular in the literature.

The Brief PEDQ-CV is a 16-item measure, with one additional item about exposure to discrimination from police. Its authors argued that the inclusion of the police item was not supported by principal components analysis, but still warrants exploratory examination, as it remains an important source of ethnicity-related stress for people of color. Items are rated on 5-point scale ranging from *never happened* (1) to *happened very often* (5). The Brief PEDQ-CV has four subscales: Stigmatization/Disvaluation,

Threat/Aggression, School/Work Discrimination, and Exclusion/Rejection, which is comparable to the original scale (Contrada et al., 2001) with one additional dimension related to school and work. Stigmatization/Disvaluation refers to experiences of being stigmatized and treated as inferior based on race (four items; e.g., implications of laziness). Threat/Aggression (four items) refers to experiences of being physically assaulted or threatened based on race. School/Work Discrimination refers to experiences of discrimination at school or work based on race (four items; e.g., unfair treatment from teachers, supervisors). Exclusion/Rejection (four items) refers to experiences of verbal harassment based on race.

The initial report of evidence of structural validity of the four-factor structure of the Brief PEDQ-CV is demonstrated using PCA. In particular, four items with the highest factor loadings on each of the subscales are retained. In partial support of discriminant validity, the scores of the Brief PEDQ-CV do not correlate, or correlate very weakly, with appraisals of discriminatory situations as challenging or beneficial. In support of convergent validity, the scores of the Brief PEDQ-CV correlate with another measure of perceived racism, the degree to which racist interactions are perceived as threatening or harmful, anxiety, cynicism, defensiveness, and the tendency to attribute hostile motivations to other people's actions. In partial support of incremental validity, the scores of the Brief PEDQ-CV correlate with appraisals of threat and harm, even when controlling for relevant personality characteristics (i.e., cynicism, defensiveness, and hostility). Internal consistency reliability estimates of the Brief PEDQ-CV subscales are adequate (above .75). Although the PEDQ-Community Version and the Brief PEDQ-CV expanded on the original scale with the use of larger Black and Latino community adult samples, its generalizability to other racial groups still remains limited in efforts of empirical cross-validation and item content not relevant for some racial groups (e.g., items across subscales related to inferiority).

Perceived Racism Scale (PRS)

The PRS (McNeilly et al., 1996a) is a 51-item comprehensive measure of perceived racism in individual, institutional, cultural, behavioral, and attitudinal domains of African American college students and community adults. Unlike other measures, the PRS comprehensively assesses (a) frequency of racism exposure (in past year and lifetime) in multiple settings, (b) emotional responses to the encounters of racism, and (c) behavioral coping responses to the encounters of racism. Participants respond to items in three separate sections of the scale. The first section of frequency is rated twice (in the past year and lifetime) on a 7-point scale ranging from *almost never* to *several times a day*. The second section of emotional response is measured by participants indicating from a number of choices how they felt during an encounter with racism and, subsequently, using a 5-point scale ranging from *not at all* to *extremely*. The third section of coping response is measured by participants choosing coping strategies they used during the encounter, ranging from active to passive responses. Frequency of Exposure Subscales include Employment (10 items), Academic (10 items), Public Realm (13 items; e.g., being followed in a store), and Racist Statements (seven items; e.g., agreeing with statements related to public assistance). Emotional Response Subscales include Anger/Frustration (eight items), Depressed Affect (16 items), and Feeling Strengthened (four items). Behavioral Coping Response Subscales include: Working Harder/Trying to Change Things (eight items), Avoiding/Ignoring (seven items), Praying (four items), Forgetting It (four items), Getting Violent (three items), and Speaking Up (four items).

The development of the PRS was guided by racism-related stress theory. It is the only measure reviewed that comprehensively assesses the frequency, emotional response, and coping behaviors of unique racialized experiences of African Americans. Evidence of structural validity of the instrument includes two PCAs of the frequency items and the behavioral/coping items based on a combined sample of 273 college student and community adults. Although, given the sample size with the actual number of items across three different domains, issues of power and factor stability are compromised. In a separate study (McNeilly et al., 1996a, 1996b), evidence of convergent validity is supported based on positive correlations with another measure of racism, cultural mistrust, and

depression. The test's authors suggest some support of discriminant validity based on insignificant relations between scores on PRS coping subscales and those of another measure of perceived racism. Internal consistency reliability estimates range from good to very good (.88–.95). Test–retest reliability estimates are generally adequate for Frequency of Exposure Subscales (.70–.80), whereas many of the Emotional Responses and Coping Responses subscales risk temporal instability (all below .70, except the Hopeless subscale).

Perceptions of Racism in Children and Youth (PRaCY)

The PRaCY (Pachter, Szalacha, Bernstein, & Garcia Coll, 2010) is a 10-item self-report measure of perceived frequency of racism among Latino, African American, West Indian/Caribbean, and multiracial children and adolescents. After each statement, participants are asked to indicate whether they experienced the type of discrimination because of the color of their skin, language or accent, or because of their culture or country of origin (i.e., yes or no), and if they had, to rate how often using a 5-point scale ranging from *once* to *weekly*. The PRaCY has two developmentally appropriate, unidimensional measures of discrimination for a younger cohort between ages 7 and 13 (10 items; e.g., receiving poor service) and an older cohort between ages 14 and 18 (10 items; e.g., unfair treatment from a police officer).

The development of PRaCY was guided by racism-related stress theory and is one of few measures that assess uniquely racialized experiences of diverse racial minority children and adolescents. However, although most items are broad enough to capture experiences across racial groups, items related to not being smart or intelligent may not be generalizable to Asian Americans. Evidence of structural validity of the instrument is based on initial examinations of interitem correlations and frequency distributions of responses with 10 items selected for each age cohort. CFA suggests excellent fit with a one-factor model for both the younger and older cohort, separately. DIF analyses further suggest no evidence of differential group bias of items based on age, sex, and ethnicity. Some evidence of convergent validity is demonstrated with significant associations between racism and depressive symptoms, physiological anxiety, and social concerns/concentration for the younger cohort, although this pattern of relationships is not consistent with the older cohort. Internal consistency reliability is adequate for both versions (.78).

Prejudice Perception Assessment Scale (PPAS)

The PPAS (Gilbert, 1998) is a five-item self-report measure based on five hypothetical vignettes of cross-racial interactions (i.e., ambiguous interpersonal situations that may be reflect racial prejudice) that assesses stigma vulnerability among African American students at predominantly White universities. After each vignette, participants respond on a 7-point scale ranging from *extremely unlikely* (1) to *extremely likely* (7) with higher scores representing greater likelihood that the interaction was attributed to racial prejudice. Two of the five vignettes reflect cross-racial, teacher–student situations. One vignette deals with a cross-racial roommate situation, and another attempts to capture cross-racial peer relationships in the classroom. A fifth vignette attempts to capture a typical collegiate activity shopping at a campus store.

The development of PPAS was guided by stigma consciousness theory and is one of few measures assessing the extent to which individuals are likely to attribute ambiguous cross-racial interactions to racial prejudice. The advantage of the vignette format avoids experimenter bias and potential ethical dilemma in exposing participants to discrimination situations, common in social psychological experiments in this area. However, generalizability of the instrument is limited to college students given the description of the vignettes. Evidence of the structural validity of the instrument is assessed using PCA on five likelihood responses, with support of a one-factor structure with all five items loading greater than .75. Evidence of convergent validity includes expected positive correlation between PPAS scores and cultural mistrust. Evidence of discriminant validity is supported by a lack of a significant correlation between PPAS scores and social

desirability. The internal consistency reliability estimate is good (.84).

Race-Related Stressor Scale (RRSS)

The RRSS (Loo et al., 2001) is a 33-item self-report measure that assesses exposure to race-related stressors in the military and combat among Asian American Vietnam war veterans. It is rated on a 5-point scale ranging from *never* to *very frequently*. RRSS has three subscales: Racial Prejudice and Stigmatization, Bicultural Identification and Conflict, and a Racist Environment. Racial Prejudice and Stigmatization (19 items) refers to direct, personal experiences of perceived discrimination, exclusion, denigration, harassment, dehumanization, or stigmatization based on race. Bicultural Identification and Conflict (seven items) refers to experiences of identifying with the Vietnamese people or culture, which conflicts psychologically with military conditioning to dehumanize the enemy. Racist Environment refers to witnessing remarks or behaviors by American military personnel that denigrated, harassed, or dehumanized Asians (seven items; e.g., fellow personnel using racial slurs).

The development of RRSS was guided by racism-related stress theory and measures uniquely racialized stressors faced by Asian American Vietnam war veterans. Evidence of structural validity of the instrument scores is assessed using EFA with support of the three-factor structure. It is worth noting intercorrelations between three subscales are large (r = .52 to .72), and questions the extent to which subscales are distinct. Evidence of convergent validity is supported with expected positive correlations between RRSS scores and combat exposure, general distress, and PTSD symptoms as well as expected negative correlations between RRSS scores and military rank. Evidence of incremental validity includes positive correlation between RRSS scores and general distress PTSD symptoms, and PTSD diagnosis, beyond effects from combat exposure and military rank. The evidence of predictive validity of RRSS scores on mental health indicators is typically large and consistent. The internal consistency reliability estimates are very good (above .90). Test–retest reliability is generally good (above .80), except the Racist Environment subscale (.69).

Scale of Ethnic Experience (SEE)

The SEE (Malcarne, Chavira, Fernandez, & Liu, 2006) is a 32-item self-report measure of acculturation and acculturative stress attitudes. It is rated on a 5-point scale ranging from *strongly disagree* (1) to *strongly agree* (5). The SEE has four subscales: Ethnic Identity, Perceived Discrimination, Mainstream Comfort, and Social Affiliation. Ethnic Identity (12 items) refers to the reflection of an individual's attitude toward being a member of an ethnic group pertaining to ethnic pride and participating in cultural activities. Perceived Discrimination refers to the individual's perceptions of how their ethnic group has been unfairly treated in the United States (nine items; e.g., perceiving that one's ethnic group is subjected to criticism). Mainstream Comfort (six items) refers to the individual's perception of comfort and identification with the "American" culture. Social Affiliation (five items) refers to the preference and comfort regarding interactions with members of their own ethnic groups.

The development of SEE was guided by acculturative stress theory, and more specifically, experiences of discrimination as part of a process of adjustment to the mainstream culture. Evidence of the structural validity of the instrument's scores is assessed using both PCA and CFA with adequate support of the four-factor structure. Evidence of convergent validity is supported by hypothesized correlations between SEE subscale scores and ethnic identity and ethnic-group specific acculturation measures. Evidence of criterion-related validity is based on expected racial group differences in SEE subscale scores. Internal consistency reliability estimates range from adequate to very good for total sample and each ethnic group (.76–.91). Although 6-week test–retest reliability coefficients for the total sample range from adequate to good (.77–.86), there is greater variation in coefficients for specific ethnic groups (.46–.86) with the lowest test–retest reliability estimate on the Perceived Discrimination subscale for African Americans. Items seem to be general enough to be meaningful across different ethnic groups (i.e., African American, White, Filipino American, and Mexican American college students). However, items do reflect both perceptions of group discrimination and personal

discrimination, which may be an issue for researchers interested in the differential effects between the two constructs.

Schedule of Racist Events (SRE)

SRE (Landrine & Klonoff, 1996) is an 18-item self-report measure of racial discrimination in different arenas (e.g., work, public places, and health care) among African American college students and community adults. After each statement, participants are asked to indicate how often they experienced the type of discrimination in their lifetime and within the past year, each on a 6-point scale ranging from *never* (1) to *almost all the time* (6). In addition, they are asked to indicate the level of stressfulness for each event on a 6-point scale ranging from *not at all stressful* (1) to *extremely stressful* (6). The three ratings are treated separately as three different subscales, Recent (Past Year), Lifetime (Entire Life), and Appraised (Stressful).

The development of SRE was guided by the racism-related stress theory and designed to assess unique negative racial discrimination experiences of African Americans. In particular, the SRE captures different sources of discrimination (e.g., teachers, neighbors), major events (e.g., being called a racial slur), daily hassles (e.g., people showing mistrust), and few behavioral and emotional reactions to discrimination (direct confrontation, anger, etc.). Although evidence of structural validity of the instrument is supported based on PCA and support of a one-factor model (Klonoff & Landrine, 1999), given item content with a mix of perceived experiences and reactions to racism, additional analyses including CFA may help clarify the factor structure of the instrument. In support of convergent validity, SRE scores generally positively correlate with general stress, another measure of discrimination, psychiatric symptoms and cigarette smoking, and inversely correlated with acculturation (e.g., Klonoff & Landrine, 1999; Klonoff, Landrine, & Ullman, 1999; Landrine & Klonoff, 1996). In support of incremental validity, SRE scores correlate positively with psychiatric symptoms, beyond socioeconomic status and generic stress (Klonoff et al., 1999). Internal consistency reliability and test–retest reliability values are very good (above .90).

Subtle and Blatant Racism Scale for Asian Americans (SABR-A^2)

The SABR-A^2 (Yoo, Steger, & Lee, 2010) is an eight-item self-report measure of subtle and blatant racism experiences among Asian American college students. It is rated on a 5-point scale ranging from *almost never* (1) to *almost always* (5). The SABR-A^2 has two subscales: Subtle Racism and Blatant Racism. Subtle Racism refers to instances of discrimination due implicitly to racial bias or stereotype (four items; i.e., treated differently, viewed with suspicion, overlooked, and faced barriers because of being Asian). Blatant Racism refers to instances of discrimination due explicitly to racial bias or stereotype (four items; i.e., called names, commented about English proficiency, physically assaulted, and made fun of because of being Asian).

The SABR-A^2 was guided by racism-related stress theory. It is the only measure reviewed assessing perceived frequency of uniquely modern, racialized experiences of Asian Americans. Both strengths and weaknesses lie in the actual number of items retained. On the one hand, an eight-item measure assessing subtle and blatant racism does not clearly capture the breadth of unique discriminatory experiences faced by Asian Americans. On the other hand, the SABR-A^2 was intentionally developed to be a brief scale that is easy to administer to stimulate research in this needed area. The evidence of structural validity of the two-factor model is strongly supported by both EFA and CFA using independent samples across two different regions of the United States. The SABR-A^2 is one of the few measures that used a separate parallel analysis (in addition to interpretability, scree plot, and eigenvalues) to determine the number of factors. Evidence of convergent validity is demonstrated with expected pattern of correlations between SABR-A^2 total and subscale scores with another measure of perceived racial discrimination, personal self-esteem, depression, anxiety, and general stress. Evidence of discriminant validity is demonstrated with expected pattern of correlations between SABR-A^2 total and subscale scores with color-blind racial attitudes. Evidence of incremental validity is partially supported with expected correlations between Blatant Racism subscale scores and

anxiety, after controlling for another measure of perceived racial discrimination. Internal consistency reliability estimates range from adequate to good (.72–.88). Two-week test–retest reliability is generally adequate (above .70), except the Blatant Racism subscale (.63).

SUMMARY AND FUTURE DIRECTIONS

The meaning and constructs of race and racism continually evolve over time and history. Race, although no longer a valid biological construct, remains an important social construct that has an effect on how individuals see themselves and others with serious implications for communal, political, and economic life. Similarly, racism continually changes its form to maintain the racial distance, power, and privilege between the racial majority and minority groups. Given the dynamic, multidimensional nature of racism, it is critical for scholars to utilize theoretically driven, psychometrically rigorous instruments capturing experiences of racism among racial minorities, thus advancing the field with accurate information on specific types, qualities, and conditions of racism that may affect psychological outcomes.

Within the past decade, psychology has seen the development of promising new measures of perceived racism with diverse theoretical models, racial populations of interest, and domains and qualities of racism assessed. Future efforts should focus on refinement and further tests of validity of these instruments. Specificity in assessment of racism experienced across racial groups or for a particular racial group should also be clarified and consistent with item development. Moreover, perceived racism measures should clarify source of discrimination. As Brondolo and colleagues (2005) noted, not all racism is perpetuated by Whites (e.g., sources of discrimination include: 53% Whites, 27% Blacks, 10% Latinos, 5% Asians, and 4% Native Americans), and there may be differential psychological effects of intergroup and intragroup racism. Finally, experiences of racism intersect with other forms of oppression (e.g., sexism and homophobia), and efforts should be made to capture these complex dynamics.

References

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Atri, A., Sharma, M., & Cottrell. (2006–2007). Role of social support, hardiness, and acculturation as predictors of mental health among international students of Asian Indian origin. *International Quarterly of Community Health Education*, *27*, 59–73. doi:10.2190/IQ.27.1.e

Bastos, J. L., Celeste, R. K., Faerstein, E., & Barros, A. J. D. (2010). Racial discrimination and health: A systematic review of scales with a focus on their psychometric properties. *Social Science and Medicine, 70*, 1091–1099. doi:10.1016/j.socscimed.2009.12.020

Belcourt-Dittloff, A., & Stewart, J. D. (2000). Historical racism: Implications for Native Americans. *American Psychologist*, *55*, 1166–1167. doi:10.1037/0003-066X.55.10.1166

Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist, 48*, 629–637. doi:10.1037/0003-066X.48.6.629

Biernat, M., & Crandall, C. S. (1999). Racial attitudes. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of political attitudes* (pp. 297–411). New York, NY: Academic Press.

Brondolo, E., Brady, N., Pencille, M., Beatty, D., & Contrada, R. J. (2009). Coping with racism: A selective review of the literature and a theoretical and methodological critique. *Journal of Behavioral Medicine, 32*, 64–88. doi:10.1007/s10865-008-9193-0

Brondolo, E., Kelly, K. P., Coakley, V., Gordon, T., Thompson, S., Levy, E., . . . Contrada, R. J. (2005). The Perceived Ethnic Discrimination Questionnaire: Development and preliminary validation of a community version. *Journal of Applied Social Psychology*, *35*, 335–365. doi:10.1111/j.1559-1816.2005.tb02124.x

Brondolo, E., Rieppi, R., Kelly, K. P., & Gerin, W. (2003). Perceived racism and blood pressure: A review of the literature and conceptual and methodological critique. *Annals of Behavioral Medicine, 25*, 55–65. doi:10.1207/S15324796ABM2501_08

Burkard, A. W., Medler, B. R., & Boticki, M. A. (2001). Prejudice and racism: Challenges and progress in measurement. In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (2nd ed., pp. 457–481). Thousand Oaks, CA: Sage.

Carter, R. T., & Pieterse, A. L. (2005). Race: A social and psychological analysis of the term and its meaning. In R. T. Carter (Ed.), *Handbook of racial–cultural psychology and counseling: Vol. 1. Theory and research* (pp. 41–63). Hoboken, NJ: Wiley.

Chavez, D. V., Moran, V. R., Reid, S. L., & Lopez, M. (1997). Acculturative stress in children: A modification of the SAFE scale. *Hispanic Journal of Behavioral Sciences, 19*, 34–44. doi:10.1177/07399863970191002

Clark, R., Coleman, A. P., & Novak, J. D. (2004). Brief report: Initial psychometric properties of the Everyday Discrimination Scale in Black adolescents. *Journal of Adolescence, 27*, 363–368. doi:10.1016/j.adolescence.2003.09.004

Contrada, R. J., Ashmore, R. D., Gary, M. L., Coups, E., Egeth, J. D., Sewell, A., . . . Chasse, V. (2001). Measures of ethnicity-related stress: Psychometric properties, ethnic group differences, and associations with well-being. *Journal of Applied Social Psychology*, *31,* 1775–1820. doi:10.1111/j.1559-1816.2001.tb00205.x

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96*, 608–630. doi:10.1037/0033-295X.96.4.608

Crocker, J., Major, B., & Steele, C. M. (1998). Social stigma. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 504–553). New York, NY: McGraw-Hill.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18. doi:10.1037/0022-3514.56.1.5

Dovidio, J. F., & Gaertner, S. L. (1998). On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. In J. L. Eberhardt & S. T. Fiske (Eds.), *Confronting racism: The problem and the response* (pp. 3–32). Thousand Oaks, CA: Sage.

Essed, P. (1991). *Understanding everyday racism.* Newbury Park, CA: Sage.

Feagin, J. R. (1991). The continuing significance of race: Anti-Black discrimination in public places. *American Sociological Review, 56*, 101–116. doi:10.2307/2095676

Feagin, J. R. (2000). *Racist America: Roots, current realities, and future reparations.* New York, NY: Routledge.

Fisher, A. B., Wallace, S. A., & Fenton, R. E. (2000). Discrimination distress during adolescence. *Journal of Youth and Adolescence*, *29,* 679–695. doi:10.1023/A:1026455906512

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 357–411). New York, NY: McGraw-Hill.

Gee, G. C., Delva, J., & Takeuchi, D. T. (2007). Relationships between self-reported unfair treatment and prescription medication use, illicit drug use, and alcohol dependence among Filipino Americans. *American Journal of Public Health, 97*, 933–940.

Gee, G. C., Ro, A., Shariff-Marco, S., & Chae, D. (2009). Racial discrimination and health among Asian Americans: Evidence, assessment, and directions for future research. *Epidemiologic Reviews, 31*, 130–151. doi:10.1093/epirev/mxp009

Gee, G. C., Spencer, M. S., Chen, J., & Takeuchi, D. (2007). A nationwide study of discrimination and chronic health conditions among Asian Americans. *American Journal of Public Health, 97*, 1275–1282. doi:10.2105/AJPH.2006.091827

Gilbert, D. J. (1998). The Prejudice Perception Assessment Scale: Measuring stigma vulnerability among African American students as predominantly Euro-American universities. *Journal of Black Psychology, 24*, 305–321. doi:10.1177/00957984980243003

Gould, S. J. (1994)., November). The geometer of race. *Discover, 15*, 64–69.

Guglielmo, J., & Salerno, S. (Eds.). (2003). *Are Italians White? How race is made in America.* New York, NY: Routledge.

Guyll, M., Matthews, K. A., & Bromberger, J. T. (2001). Discrimination and unfair treatment: Relationship to cardiovascular reactivity among African American and European American women. *Health Psychology, 20*, 315–325. doi:10.1037/0278-6133.20.5.315

Harrell, S. P. (2000). A multidimensional conceptualization of racism-related stress: Implications for the well-being of people of color. *American Journal of Orthopsychiatry, 70*, 42–57. doi:10.1037/h0087722

Helms, J. E. (1990). The measurement of racial identity attitudes. In J. E. Helms (Ed.), *Black and White racial identity: Theory, research and practice* (pp. 33–48). New York, NY: Greenwood Press.

Helms, J. E., & Cook, D. A. (1999). *Using race and culture in counseling and psychotherapy: Theory and process.* Needham Heights, MA: Allyn & Bacon.

Helms, J. E., & Talleyrand, R. (1997). Race is not ethnicity. *American Psychologist, 52*, 1246–1247. doi:10.1037/0003-066X.52.11.1246

Jones, J. M. (1997). *Prejudice and racism* (2nd ed.). New York, NY: McGraw-Hill.

Karlsen, S., & Nazroo, J. Y. (2002). Relation between racial discrimination, social class, and health among ethnic minority groups. *American Journal of Public Health, 92*, 624–631. doi:10.2105/AJPH.92.4.624

Kessler, R. C., Mickelson, K. D., & Williams, D. R. (1999). The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. *Journal of Health and Social Behavior, 40*, 208–230. doi:10.2307/2676349

Klonoff, E. A., & Landrine, H. (1999). Cross-validation of the Schedule of Racist Events. *Journal of Black Psychology, 25*, 231–254. doi:10.1177/0095798499025002006

Klonoff, E. A., Landrine, H., & Ullman, J. B. (1999). Racial discrimination and psychiatric symptoms among Blacks. *Cultural Diversity and Ethnic Minority Psychology, 5*, 329–339. doi:10.1037/1099-9809.5.4.329

Kressin, N. R., Raymond, K., & Manze, M. (2008). A review of measures of perceived race/ethnic-Based discrimination in health care. *Journal of Health Care for the Poor and Underserved, 19*, 697–730. doi:10.1353/hpu.0.0041

Krieger, N., & Sidney, S. (1996). Racial discrimination and blood pressure: The CARDIA study of young black and white adults. *American Journal of Public Health, 86*, 1370–1378. doi:10.2105/AJPH.86.10.1370

Krieger, N., Smith, K., Naishadham, D., Hartman, C., & Barbeau, E. M. (2005). Experiences of discrimination: Validity and reliability of a self-report measure for population health research on racism and health. *Social Science and Medicine, 61*, 1576–1596. doi:10.1016/j.socscimed.2005.03.006

Kwan, K. L. K. (2005). Racial salience: Conceptual dimensions and implications for racial and ethnic identity development. In R. T. Carter (Ed.), *Handbook of racial–cultural psychology and counseling: Vol. 1. Theory and research* (pp. 115–131). Hoboken, NJ: Wiley.

Landrine, H., & Klonoff, E. A. (1996). The Schedule of Racist Events: A measure of racial discrimination and a study of its negative physical and mental health consequences. *Journal of Black Psychology*, *22,* 144–168. doi:10.1177/00957984960222002

Landrine, H., Klonoff, E. A., Corral, I., Fernandez, S., & Roesch, S. (2006). Conceptualizing and Measuring ethnic discrimination in health research. *Journal of Behavioral Medicine*, *29,* 79–94. doi:10.1007/s10865-005-9029-0

Liang, C. T. H., & Fassinger, R. E. (2008). The role of collective self-esteem for Asian Americans experiencing racism-related stress: A test of moderator and mediator hypotheses. *Cultural Diversity and Ethnic Minority Psychology, 14*, 19–28. doi:10.1037/1099-9809.14.1.19

Liang, C. T. H., Li, L. C., & Kim, B. S. K. (2004). The Asian American Racism-Related Stress Inventory: Development, factor analysis, reliability, and validity. *Journal of Counseling Psychology*, *51,* 103–114. doi:10.1037/0022-0167.51.1.103

Loo, C. M., Fairbank, J. A., Scurfield, R. M., Ruch, L. O., King, D. W., Adams, L. J., & Chemtob, C. M. (2001). Measuring exposure to racism: Development and validation of a Race-Related Stressor Scale (RRSS) for Asian American Vietnam veterans. *Psychological Assessment*, *13*, 503–520. doi:10.1037/1040-3590.13.4.503

Lopez, I. (2006). *White by law 10th anniversary edition: The legal construction of race*. New York, NY: New York University Press.

Malcarne, V. L., Chavira, D. A., Fernandez, S., & Liu, P. (2006). The Scale of Ethnic Experience: Development and psychometric properties. *Journal of Personality Assessment*, *86,* 150–161. doi:10.1207/s15327752jpa8602_04

McNeilly, M. D., Anderson, N. B., Armstead, C. A., Clark, A. R., Corbett, M., Robinson, E. L., . . . Lepisto, E. M. (1996a). The Perceived Racism Scale: A multidimensional assessment of the experience of White racism among African Americans. *Ethnicity and Disease*, *6*, 154–166.

McNeilly, M. D., Anderson, N. B., Robinson, E. L., McManus, C., & Armstead, C. A., Clark, R., et al. (1996b). The convergent, discriminant, and concurrent criterion validity of the Perceived Racism Scale: A multidimensional assessment of the experience of White racism among African Americans. In R. L. Jones (Ed.), *Handbook of tests and measurements for Black populations* (Vol. 2, pp. 359–374). Hampton, VA: Cobb & Henry.

Miller, J., & Garran, A. M. (2008). *Racism in the United States: Implications for the helping professions*. Belmont, CA: Brooks/Cole.

Min, P. G. (Ed.). (2005). *Encyclopedia of racism in the United States*. Westport, CT: Greenwood Press.

National Research Council. (2004). *Measuring racial discrimination*. Washington, DC: National Academies Press.

Neville, H. A., Lilly, R. L., Duran, G., Lee, R. M., & Browne, L. (2000). Construction and initial validation of the Color-Blind Racial Attitude Scale (CoBRAS). *Journal of Counseling Psychology*, *47,* 59–70. doi:10.1037/0022-0167.47.1.59

Noh, S., & Kaspar, V. (2003). Perceived discrimination and depression: Moderating effects of coping, acculturation, and ethnic support. *American Journal of Public Health, 93*, 232–238. doi:10.2105/AJPH.93.2.232

Omi, M., & Winant, H. (1994). *Racial formation in the United States: From the 1960s to the 1990s* (2nd ed.). New York, NY: Routledge.

Pachter, L. M., Szalacha, L. A., Bernstein, B. A., & Garcia Coll, C. (2010). Perceptions of Racism in Children and Youth (PRaCY): Properties of a self-report instrument for research on children's health and development. *Ethnicity and Health, 15*, 33–46. doi:10.1080/13557850903383196

Paradies, Y. (2006). A systematic review of empirical research on self-reported racism and health. *International Journal of Epidemiology, 35*, 888–901. doi:10.1093/ije/dyl056

Plous, S. (2003). The psychology of prejudice, stereotyping, and discrimination: An overview. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 3–48). New York, NY: McGraw-Hill.

Racism. In *Merriam-Webster's collegiate dictionary* (11th ed.). (2003). Springfield, MA: Merriam-Webster.

Richards, G. (1997). *Race, racism, and psychology: A reflexive history*. London, England: Routledge.

Root, M. P. P. (1993). Reconstructing the impact of trauma on personality. In M. Ballou & L. Brown (Eds.), *Theories of personality and psychopathology: Feminist reappraisal* (pp. 229–265). New York, NY: Guilford Press.

Root, M. P. P. (2000). Rethinking racial identity development. In P. Spickard & W. J. Burroughs (Eds.), *We are a people: Narrative and multiplicity in constructing ethnic identity* (pp. 205–220). Philadelphia, PA: Temple University Press.

Sam, D. L., & Berry, J. W. (Eds.). (2006). *The Cambridge handbook of acculturation psychology*. Cambridge, England: Cambridge University Press.

Saucier, D. A., Miller, C. T., & Doucet, N. (2005). Differences in helping Whites and Blacks: A meta-analysis. *Personality and Social Psychology Review, 9*, 2–16. doi:10.1207/s15327957pspr0901_1

Seaton, E. K. (2003). An examination of the factor structure of the Index of Race Related Stress among a sample of African American Adolescents. *Journal of Black Psychology, 29*, 292–307. doi:10.1177/0095798403254211

Seaton, E. K. (2006). Examination of a measure of racial discrimination among African American adolescents. *Journal of Applied Social Psychology*, *36*, 1414–1429. doi:10.1111/j.0021-9029.2006.00066.x

Smedley, A., & Smedley, B. (2005). Race as biology is fiction, racism as a social problem is real. *American Psychologist, 60*, 16–26. doi:10.1037/0003-066X.60.1.16

Smedley, B. D., Myers, H. F., & Harrell, S. P. (1993). Minority-status stresses and the college adjustment for ethnic minority freshmen. *Journal of Higher Education*, *64,* 434–452. doi:10.2307/2960051

Sodowsky, G. R., & Plake, B. S. (1991a). Moderating effects of sociocultural variables on acculturation attitudes of Hispanics and Asian Americans. *Journal of Counseling and Development, 70*, 194–204. doi:10.1002/j.1556-6676.1991.tb01583.x

Sodowsky, G. R., & Plake, B. S. (1991b). Psychometric properties of the American-International Relations Scale. *Educational and Psychological Measurement, 51*, 201–216.

Sodowsky, G. R., & Plake, B. (1992). A study of acculturation differences among international people and suggestions for sensitivity to within-group differences. *Journal of Counseling and Development, 71*, 53–59. doi:10.1002/j.1556-6676.1992.tb02171.x

Stancil, T. R., Hertz-Picciotto, I., Schramm, M., & Watt-Morse, M. (2000). Stress and pregnancy among African–American women. *Paediatric and Perinatal Epidemiology, 14*, 127–135. doi:10.1046/j.1365-3016.2000.00257.x

Stuber, J., Galea, S., Ahern, J., Blaney, S., & Fuller, C. (2003). The association between multiple domains of discrimination and self-assessed health: A multilevel analysis of Latinos and blacks in four low-income New York City neighborhoods. *Health Services Research*, *38*, 1735–1760. doi:10.1111/j.1475-6773.2003.00200.x

Suarez-Morales, L., Dillon, F. R., & Szapocznik, J. (2007). Validation of the Acculturative Stress Inventory for Children. *Cultural Diversity and Ethnic Minority Psychology, 13*, 216–224. doi:10.1037/1099-9809.13.3.216

Sue, D. W. (2005). Racism and the conspiracy of silence: Presidential address. *The Counseling Psychologist, 33*, 100–114. doi:10.1177/0011000004270686

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62*, 271–286. doi:10.1037/0003-066X.62.4.271

Swim, J. K., & Stangor, C. (1998). *Prejudice: The target's perspective.* New York, NY: Academic Press.

Takaki, R. (1993). *A different mirror: A history of multicultural America.* Boston, MA: Little, Brown.

Tatum, B. D. (1997). *Why are all the Black kids sitting together in the cafeteria? And other conversations about race*. New York, NY: Basic Books.

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, Public Health Services, Office of the Surgeon General.

Utsey, S. (1999). Development and validation of the Index of Race-Related Stress (IRRS)—Brief version. *Measurement and Evaluation in Counseling and Development, 32*, 149–167.

Utsey, S. O., Chae, M. H., Brown, C. F., & Kelly, D. (2002). Effect of ethnic group membership on ethnic identity, race-related stress, and quality of life. *Cultural Diversity and Ethnic Minority Psychology, 8*, 366–377. doi:10.1037/1099-9809.8.4.367

Utsey, S. O., & Ponterotto, J. G. (1996). Development and validation of the Index of Race-Related Stress (IRRS). *Journal of Counseling Psychology*, *43,* 490–501. doi:10.1037/0022-0167.43.4.490

Webster's unabridged dictionary (2nd ed.). (2005). New York, NY: Random House.

Whitbeck, L. B., Adams, G. W., Hoyt, D. R., & Chen, X. (2004). Conceptualizing and measuring historical trauma among American Indian people. *American Journal of Psychology, 33*, 119–130. doi:10.1023/B:AJCP.0000027000.77357.31

Williams, D. R., & Mohammed, S. A. (2009). Discrimination and racial disparities in health: Evidence and needed research. *Journal of Behavioral Medicine, 32*, 20–47. doi:10.1007/s10865-008-9185-0

Williams, D. R., Yu, Y., Jackson, J. S., & Anderson, N. B. (1997). Racial differences in physical and mental health: Socio-economic status, stress, and discrimination. *Journal of Health Psychology, 2*, 335–351. doi:10.1177/135910539700200305

Winston, A. S. (Ed.). (2004). *Defining difference: Race and racism in the history of psychology*. Washington, DC: American Psychological Association. doi:10.1037/10625-000

Wu, F. H. (2002). *Yellow: Race in America beyond Black and White*. New York, NY: Basic Books.

Yoo, H. C., Burrola, K. S., & Steger, M. F. (2010). A preliminary report on a new measure: Internalization of the Model Minority Myth Measure (IM-4) and its psychological correlates among Asian American college students. *Journal of Counseling Psychology*, *57*, 114–127. doi:10.1037/a0017871

Yoo, H. C., Steger, M. F., & Lee, R. M. (2010). Validation of the Subtle and Blatant Racism Scale for Asian American college students (SABR-A^2). *Cultural Diversity and Ethnic Minority Psychology*, *16*, 323–334. doi:10.1037/a0018674

CHAPTER 26

THERAPEUTIC ASSESSMENT: USING PSYCHOLOGICAL TESTING AS BRIEF THERAPY

Stephen E. Finn and Hale Martin

The field of psychological assessment is undergoing dramatic change, perhaps even a paradigm shift. Traditionally, assessment has focused on gathering accurate data to use in clarifying diagnoses and developing treatment plans. Although largely retaining these goals, new approaches also emphasize the therapeutic effect assessment can have on clients and important others in their lives. Evidence accumulating over the past 20 years suggests that this effect can be substantial. On the basis of their meta-analysis of outcome studies investigating the therapeutic effect that assessment can have, Poston and Hanson (2010) contended that psychology needs to reconsider training in assessment to incorporate approaches that emphasize its therapeutic value. They even argued that managed care organizations need to reevaluate delivery of services in light of the efficacy of the new approaches to assessment. This chapter highlights the development of the therapeutic application of psychological assessment, examines its empirical support, discusses how assessment might produce therapeutic change, and outlines the Therapeutic Assessment (TA) approach and illustrates it through a case example.

WHAT IS TA?

TA is a semistructured approach to assessment that strives to maximize the likelihood of therapeutic change for the client. It has been developed largely through the efforts of Stephen Finn and his colleagues (Finn, 1996, 2007; Finn & Martin, 1997; Finn & Tonsager, 1997), building on the innovations of collaborative assessment developed by Constance Fischer (1985/1994), Leonard Handler (2006); Caroline Purves (2002), and others. TA has incorporated knowledge from a range of psychology to produce an evidence-based approach to positive personal change through psychological assessment. It rests on the commonsense application of the powerful insights that are efficiently available through reliable and valid assessment tools and techniques to a collaborative, respectful, supportive, gentle, and ultimately experiential process of self-discovery.

DEVELOPMENT OF TA

The roots of TA are grounded in the early work of Constance Fisher. In 1978, she wrote that historically "psychology has assumed that people should be treated as objects amenable to measurement, prediction, and control" (p. 41), and argued that psychologists do not have to be restricted this way but rather can acknowledge that humans are purposeful and that the "the professional's understandings are not more real, valid, or influential within the client's life than are the client's" (p. 42). Fischer then defined collaborative assessment as assessment "in which the client and professional labor together toward mutually set goals, sharing their respective background information and emerging impression" (p. 42).

Fischer's work largely provoked resistance until Finn discovered it and saw that it fit nicely into his own understanding of the value of assessment. The collaborative/TA movement began in earnest in 1992 when Finn and his graduate

DOI: 10.1037/14048-026
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

student, Mary Tonsager, published a randomized controlled study showing that a simple assessment involving only the Minnesota Multiphasic Personality Inventory—2 (MMPI–2)—but conducted in a collaborative manner—had powerful therapeutic benefits. Over succeeding years, Finn has developed the TA approach, incorporating evolving insights from other areas of psychology, including the idea that emotional experience rather than logical understanding is at the heart of therapeutic change (Fosha, 2000; Schore, 2009). As the case example in this chapter illustrates, TA involves structured experiential components that move assessment from an intellectual exercise to an experiential one.

Through these developments, TA has been applied to a range of assessment situations. Finn adapted the approach developed for adults to the assessment of children. He and his colleagues borrowed a practice from Fischer (1985/1994) and others and introduced the idea of writing fables or stories as age-appropriate feedback for children, offering new options and outcomes through stories (Tharinger, Finn, Wilkinson, et al., 2008). More recently, Deborah Tharinger, Finn, and their students at the University of Texas at Austin have further advanced the assessment of children to integrate the child's parents and support system into the process (Tharinger, Finn, Austin, et al., 2008; Tharinger, Finn, Hersh, et al., 2008). This same group is also researching applications of TA to adolescents (Tharinger, Finn, Gentry, & Matson, in press). Finn also pioneered the application of TA to couples' assessments; these often involve using a consensus Rorschach to help partners understand and undo projective identification (Finn, 2007). Others are applying TA to neuropsychological assessment (Gorske & Smith, 2008).

WHAT IS THE EVIDENCE BASE FOR TA?

As mentioned earlier, Finn and Tonsager (1992) were the first to document empirically the therapeutic effect an assessment can have. They studied students waiting for therapy at a university counseling center and randomly assigned them to two conditions. About half ($n = 32$) underwent a two-session assessment in which they took the MMPI–2 and were given feedback according to a collaborative method developed by Finn (1996). The remaining students ($n = 29$) got equal clinical attention focused on their current concerns but no assessment. The effects were striking. The assessment group showed a significant drop in symptoms ($p < .01$; effect size = .85) and a significant increase in self-esteem ($p < .001$; effect size = .46), whereas both these variables remained unchanged in the nonassessment group across time. It is interesting that the two groups showed no differences in feelings about the therapist/assessor. This simple study provided powerful evidence that assessment can be therapeutic, and it fueled the movement to use assessment therapeutically.

The Finn and Tonsager (1992) research was replicated by Newman and Greenway (1997), who conducted a similar study with some refinements. In their investigation, both groups took the MMPI–2, although the control group received feedback after the outcome measures were completed. Their results were similar to those in the Finn and Tonsager study, but the effect sizes were smaller. Subsequently, a growing number of studies have investigated various effects of TA. The research that has been done is intriguing and largely supports the contention that TA is an effective therapeutic intervention. Ackerman, Hilsenroth, Baity, and Blagys (2000) showed that TA was better than traditional assessment in insuring compliance with treatment recommendations (effect size = .42). Ougrin, Ng, and Low (2008) found that TA was superior to "assessment as usual" in getting adolescents admitted to emergency rooms for self-harm to attend follow-up appointments ($p < .05$) and engage with services ($p < .05$). Hilsenroth, Peters, and Ackerman (2004) demonstrated that TA was better than traditional information gathering assessment in strengthening clients' therapeutic alliance with a subsequent psychotherapist (effect size = 1.02).

Little and Smith (2008) studied psychiatric inpatients, showing that TA was more effective than structured supportive therapy or standard psychiatric treatment and milieu therapy on several factors, including facilitating treatment alliance, cooperation with treatment, and satisfaction with treatment, and

that it also promoted lower distress and an increased sense of well-being. Morey, Lowmaster, and Hopwood (2010) showed that a brief (two-session) TA improved therapeutic outcomes in a group of women diagnosed with borderline personality disorder who were undergoing a brief, manualized form of cognitive therapy. Tharinger, Finn, Gentry, Hamilton, et al. (2009) showed decreased symptomatology in both latency-aged children and their mothers, decreased family conflict, and better family communication after a nine-session TA. Finally, a study by Smith, Handler, and Nash (2010) used a time-series analysis to reveal improvement in symptomatology and family relationships in latency-aged males with oppositional defiant disorder who underwent a TA with their caregivers.

The most telling study of collaborative/TA is a meta-analysis by Poston and Hanson (2010). They identified 17 published studies (with a total 1,496 participants) that used psychological assessment as a therapeutic intervention. They defined therapeutic intervention "broadly as the process of completing any formal psychological test/measure and receiving feedback on the results . . . with therapeutic intent" (p. 205). Thus, many of the studies they included only examined the therapeutic effects of feedback.

The results showed an overall effect size of .423, which was significant at the .01 level. Poston and Hanson (2010) compared this with the effect sizes of various psychotherapy approaches and found it comparable with substance abuse treatment (effect size = .45) and "approaching" cognitive–behavioral treatment for anxiety disorders (effect size ranged from .89 to 2.59) and general psychotherapy (effect size = .80). This is an impressive showing for TA, given that some of the assessments in the study involved as few as two sessions.

From these results, Poston and Hanson (2010) concluded that "those who engage in assessment and testing as usual may miss out, it seems, on a golden opportunity to effect client change and enhance clinically important treatment processes" (p. 210). Furthermore, they asserted that (a) training programs should include therapeutic assessment models, (b) competency benchmarks should include aspects of TA, and (c) managed care managers should consider TA in future policies.

HOW CAN SUCH A BRIEF INTERVENTION BE EFFECTIVE?

There are several ingredients of TA that may help explain why such a brief intervention can have powerful therapeutic effects. Different therapeutic elements fit nicely into the assessment situation and in that context likely potentiate change.

Changing a Client's Self-Narrative

TA changes the narrative clients have developed about themselves or about their children or spouses. Effectively changing how people view themselves in the world opens new possibilities in their lives. For example, certain clients who have always thought of themselves as "stupid" may discover through a TA that they are intelligent but have a specific learning disability that impeded them in school. Learning this information may dramatically change the way the clients view themselves and the choices they make after the TA.

Walking the Line Between Self-Verification and Disintegration

One of the challenges in creating change in any therapeutic process is walking the line between self-verification and the potential for disintegration. Self-verification is the powerful human tendency to seek and attend to information that supports established ways one understands oneself (Swann, 1997). Even when this view is negative and self-limiting, a person will hold on to it in the face of more favorable understandings (Swann, Wenzlaff, Krull, & Pelham, 1992). Practicing psychologists have historically labeled such attachment to old ways of thinking as "resistance." Research in social psychology has helped determine that reliance on established self-views serves to "predict the reactions of others, to guide behavior, and to organize one's conceptions of reality" (Swann, 1997, p. 177). If these established patterns of understanding oneself in the world are changed too abruptly, the person risks feelings of disintegration, an experience of emotional distress, disorientation, and fear that can result when an individual is unable to refute evidence that some central and tightly held belief about the self is wrong (Kohut, 1984). The balance between self-verification

and disintegration is central in promoting change. The focus on empathy and relationship in TA—as well as specific TA techniques such as involving clients in setting the goals for an assessment and in interpreting test results—enables the assessor to facilitate change without overwhelming the client.

Using Psychological Tests as "Empathy Magnifiers"

Another aspect of psychological change that TA harnesses is the power of empathy. The client is changed by the experience of being deeply seen and understood (Kohut 1982). Psychological tests are excellent "empathy magnifiers" (Finn, 2007) and, thus, perfectly suited to maximize the empathy experience. Using empathic understanding gently and with compassion creates a unique experience of being seen that is healing in itself for many clients (Finn, 2009).

Involving the Entire System With Children and Families

TA is a family-systems intervention with children and families. Thus, caretakers observe or participate in each step in the assessment process and are included in discussions during or after each assessment session. During these discussions, the parents are led by the data and their interactions to see their child more accurately and understand what the child needs. The assessor also helps parents deal with their own pain and limitations. Thus, TA addresses the child's most important interpersonal environment to create opportunity for growth and therapeutic change.

Undoing Projective Identification With Couples

The challenge in couples treatment is often exposing and diminishing the relationship patterns rooted in childhood that guide expectations, reactions, and behavior with one's spouse or partner. These patterns are often deeply entrenched and completely ego-syntonic. TA approaches couples work by first assessing the partners independently to gain an understanding of each person's underlying dynamics that influence the couple's relationship. By beginning to understand how each partner imposes his or her relationship history on the other, the assessor can move the couple to relate more from accurate, present-day reality than from projection of past relationship patterns, which are often shaded with hurts and failures. Often, when each spouse begins to understand the reasons why the other acts in certain ways, repair and healing can begin.

STEPS IN THE TA PROCESS ILLUSTRATED THROUGH AN ADULT CASE

Finn and Tonsager (1997) articulated the semistructured approach of TA in sequential steps, and these were expanded by Finn (2007). In this section, we briefly discuss each step and use a case example to illustrate each of the steps in practice.

Initial Contact

The assessment begins with the initial contact with the referring professional and later with the client. Usually these both happen by telephone. Questions and information are sought from the referring professional, and he or she is encouraged to share the questions with the client.

The initial phone contact with clients and sometimes even the recorded message encountered convey a wealth of information: how they present themselves, what concerns they might have, their tone of voice, and how open they are to the assessment. A collaborative assessor–client relationship is begun in the initial phone contact by asking clients to think of questions they would like the assessment to answer. The assessor also answers practical questions and schedules the first meeting.

Initial Contact: Case Illustration

A well-respected therapist, Sarah, contacted Steve for a TA to aid in outpatient therapy with Luanne, a 26-year-old woman who had been working with Sarah in therapy for about 9 months. Sarah explained that Luanne was having trouble "settling in" to therapy because of her loyalty to her previous therapist, Mary, whom she had seen for 6 years. Luanne had begun therapy with Sarah immediately after relocating for school. Sarah reported that Luanne still talked to Mary several times a week by phone. The focus of Luanne's treatment was her

recovery from childhood sexual abuse by her father. This abuse had gone on for many years, and Luanne's mother had apparently known about or suspected the abuse but had done nothing.

Sarah explained that in Luanne's previous treatment, Luanne had done a great deal of emotionally intense "reliving" of past traumas and psychodrama enactments of confrontations with family members. Sarah felt that such a therapeutic approach was not what Luanne needed currently, and she wanted to work with Luanne on experiencing her emotions in a "modulated way" that was less disruptive to her life. Sarah said that Luanne perceived this approach as a message of "You have to stuff your feelings." Sarah had two questions she wanted the assessment to address: "What will help Luanne shift her alliance more from her previous therapist to me?" and "How can I avoid getting in a power struggle with her about how to work on feelings in therapy?"

Initial Session

The initial session is very important as it sets the frame in which the assessment will occur. The assessor tries to convey authentic warmth, respect, compassion, and curiosity and to engage the client as a collaborator. Sometimes clients are taken aback by the expectation that their questions will drive the assessment and that they will play an active role. They may need help understanding that psychological tests are not "oracles" and that their coparticipation is essential for the assessment to be valid and useful. Sometimes clients need help formulating questions. If so, to the assessor can encourage them to talk about the problems they are having in life and then listen carefully for potential questions to bring to their attention.

As questions that the client sees as central come into focus, the assessor gathers relevant background for each. It is also helpful to inquire about past assessments and any hurts they might have caused. This can be an important step in insuring that the assessor and the client will not repeat those injuries. In TA, the assessor also asks clients if they have questions about the assessor. This simple act conveys that the relationship is open both ways. Rarely do clients ask anything inappropriate, but they do have a chance to address any concerns or fears they have about the assessor or the assessment. Before parting, the assessor and client review the client's questions and the plan of work and agree on fees and the schedule of future sessions.

Initial Session: Case Illustration

Luanne was a tall, handsome, athletic-looking woman, who greeted Steve in the waiting room with direct eye contact and a firm handshake. Luanne seemed very comfortable and fairly quickly articulated her first question for the assessment: "Is there a way for me to not be as controlled as I am by shame?" Luanne explained that she felt held back in many situations—with friends, in school, in social situations—by the fear that she would do something "wrong" and "look like a fool." She said she had always had intense shame but that it was worse in her relationships with men, where she rarely spoke up and had a very difficult time "holding on to herself." She explained that she had not dated for 7 years.

This led Luanne to pose a second question: "What still gets in the way of my dating?" She briefly mentioned her childhood sexual abuse and said that as an adolescent and adult, she had dated abusive men who "treated her like dirt." Luanne said that currently, she was longing to start dating again but was also "terrified" that she would get back into old patterns. She and Steve agreed to see what the testing could help her understand about her fear.

Steve then asked Luanne if she and Sarah had a game plan for working on the dating, and Luanne finally began talking about her ambivalence regarding the therapy with Sarah. Luanne explained that the therapy she had done with Mary was very different. The philosophy had been "all feelings should be felt," and "if it doesn't hurt, you're not working." Luanne said Sarah's goal of helping her develop affect regulation seemed like a "waste of time" and that she feared she wasn't going to "get all her feelings out" so she could go on and lead a normal life. Steve asked Luanne if she felt she was making changes in her life as a result of her work with Sarah. She said she was and that actually she was functioning better than she had in years. This seemed to surprise her and led Luanne to pose her first question about the therapy: "What's the best approach for me in therapy: pushing for lots of

feeling versus a more paced, controlled approach?" When Steve asked, "Does it have to be either/or?" Luanne admitted that she tended to think in "black-and-white terms" and asked another question: "Is it possible and/or desirable to integrate these two approaches to treatment?"

Toward the end of the meeting, Steve and Luanne talked about practical aspects of the assessment. Her funds were limited, so they agreed to use the MMPI–2 (Butcher et al, 1989) as the main assessment instrument and made arrangements for Luanne to take the test at Steve's office before their next meeting. Steve asked Luanne what it had been like to talk together that day, and Luanne said, "More comfortable than I thought it would be. I've never had a male therapist before. But you were easy to talk to, and I feel excited about doing the testing together."

After Luanne left, Steve was aware of feeling somewhat sad. He wondered if Luanne needed to understand that even after successful treatment, her sexual abuse would always play some role in her life. It seemed possible that part of Luanne's dilemma about therapy was her fantasy that if she just worked on her trauma enough and "got her feelings out," it would be as if she were "washed clean" and it had never happened. Steve knew this was not possible and wondered if her previous therapy had reinforced Luanne's fantasy.

Steve also noticed how much Luanne seemed to look to external sources to guide her decision making. Luanne had talked as if she had little choice about the pacing of her therapy rather than exploring her own mixed feelings about going fast or slow. Steve decided to pay special attention to helping Luanne make her own choices during the assessment, remembering that she said she tended to "give herself away" in relationships, especially with men.

Standardized Testing Sessions

Testing typically begins at the next session. Tests are administered in standardized ways to gather information that will inform the answers to the questions. To begin, the assessor often chooses tests that are more clearly related to the client's questions. This conveys that the assessor is indeed focusing on issues the client has identified. One technique that has become increasingly valued in TA is the extended inquiry (Handler, 1999). This technique involves the assessor asking about the client's experience of a test or the client's thoughts about certain test responses.

Standardized Testing Sessions: Case Illustration

Luanne's MMPI–2 showed no signs of invalidity, and it appeared that she approached the test in a very unguarded manner (Variable Response Inconsistency [VRIN] = 46T; True Response Inconsistency [TRIN] = 65T; Infrequency Psychopathology [Fp] = 49T; Lie [L] = 42T; and Defensiveness [K] = 37T). This kind of openness is not uncommon in clients voluntarily taking part in TA who have defined personal questions they want to have answered using the MMPI–2. The moderate elevation on F (Infrequency; 79T) was higher than that found in most outpatient therapy clients and indicated that Luanne was in a significant amount of distress, more than Steve had picked up on in the initial interview.

This distress was confirmed by the profile of clinical scales, where Luanne had seven scales with significant elevations: Scale 1 (Hypochondriasis; 69T), Scale 2 (Depression; 68T), Scale 4 (Psychopathic Deviate; 85T), Scale 6 (Paranoia; 74T), Scale 7 (Psychasthenia; 76T), Scale 8 (Schizophrenia; 77T), and Scale 0 (Social Introversion; 77T). Scale 3 (Hysteria; 56T) was slightly elevated, while Scale 5 (Masculinity-Femininity; 45T), and Scale 9 (Mania; 54T) were not elevated. This "gull-wing" configuration is not unusual among women with histories of trauma and current difficulties with relationships (Graham, 2006). According to Caldwell's (2001) theory, Luanne's profile suggested that she was a "sturdy survivor" with a traumatic childhood who had been exposed to humiliating and shocking events without adequate support and who had coped by "pulling herself up by her "bootstraps," focusing on achievement and avoiding intimacy. Steve hypothesized that this coping strategy was one reason Luanne did not seem as distressed in person as she appeared on the MMPI–2.

Apart from distress, the MMPI–2 profile indicated problems in a number of areas. Women with similar profiles have identity confusion, histories of

drug and alcohol abuse, and tumultuous relationships. They have problems with emotion regulation and emotional flooding, tend to engage in splitting and "black-and-white thinking," and can go through periods—particularly when emotionally aroused—when their thinking is illogical and distorted, especially in the area of interpersonal relationships. Steve was surprised by the elevation on Scale 0 (Social Introversion) because it did not fit his picture of Luanne having been deeply embedded in her previous therapeutic community. Further examination of the subscales for Scale 0 (Ben-Porath, Hostetler, et al., 1989) showed that Luanne was a "sociable introvert"; that is, she desired contact with other people but tended to avoid social situations and relationships because of her anxiety and low self-esteem. Regarding low-self esteem and shame, Luanne had a high score (84T) on the Low Self-Esteem scale of the MMPI–2, suggesting that she was self-critical and often felt worthless and insignificant.

On the basis of this information, Steve believed that he had an understanding of Luanne's struggle with shame (her first question), that there were many good reasons why she was avoiding dating (her second question), and that a more paced therapeutic approach was likely to be the most beneficial (her third question). In general, however, Steve wanted to avoid imposing this understanding on Luanne, with the hope that Luanne could take more of her own authority in choosing how to pace her treatment. If Luanne could feel more in charge, it might make her feel less afraid of the world. With these goals in mind, Steve planned an assessment intervention.

Assessment Intervention Session

Perhaps the most innovative step in TA is the assessment intervention session. In this session, the assessor uses the information gathered up to that point to elicit an analogue of the client's main difficulties in vivo. If this is successful, the assessor invites the client to observe the problem behavior, understand it, and then solve it in the assessment session. Meanwhile, the assessor and client relate their discussions to the client's daily life. The assessment intervention session is a stepping stone to answers that will be discussed in the upcoming summary/discussion session.

Assessment Intervention Session: Case Illustration

In short, Steve's plan for an assessment intervention was to arouse Luanne emotionally in a controlled fashion while keeping close tabs on her level of distress and to put her in the driver's seat about whether she wanted to "push for more feelings" or "slow things down." In preparation, Steve selected cards from a number of picture story tests and ordered them according to his sense of their emotional difficulty for Luanne. Steve introduced the session to Luanne as follows: "Today I want to do another test with you that I hope will help us explore your question, 'What's the best approach for me in therapy: pushing for lots of feeling vs. a more paced, controlled approach?'" Steve explained that the test they would be working with might be emotionally arousing and asked if Luanne was OK with that. She said she was, and Steve then gave the standard instructions for the Thematic Apperception Test (TAT; Murray, 1943).

In order, Steve then asked Luanne to tell stories to pictures of a woman sitting, resting on the back of chair while looking off into the distance (TAT card 8GF); of a young teenage girl sitting on a curb in front of a house looking at her hands (Card 2F of the Adolescent Apperception Cards; Silverton, 1993); and of an androgynous-looking teen sitting up in bed under the covers while an adult man sits at the foot of the bed with his hand on the teen's thigh (Card 2 of the Family Apperception Test; Sotile, Henry, & Sotile, 1988). Luanne easily told stories to the first two cards, and each contained themes of the characters being "bored and lonely." Steve chose the third card because it could suggest sexual abuse. Luanne told a story of a father tucking in his child who, for some reason, did not "feel safe" because "it's hard to predict how the dad is going to be at any moment." She said the child "wished the dad would go away." There was no explicit mention of sexual abuse or violence, but Luanne looked quite uncomfortable as she told the story. Afterward, Steve asked how Luanne was doing, and she said, "Fine. I can do more." He then presented her with

TAT Card 13MF (a woman lying on a bed with a man standing nearby), and Luanne told the following story:

> Wow. I thought I didn't like the lonely female pictures, but this is the worst. This is a couple and she looks passed out, like she's spent a lot of time that night running from herself and from reality and is out of it. Her husband or boyfriend is tired of seeing this again. He has come home from work and can't stand seeing this again. He doesn't think about why she did this, just about how hard it is to deal with. He doesn't cover her up, he covers his eyes and thinks about himself. He'll probably go have a drink afterwards and not realize he's doing just what she does. [Steve: How does she run away from reality?] Through drugs and alcohol. It's really a shame. She looks really hungry for love and he's not available for it.

Afterward, Luanne looked somewhat "blank" to Steve, so he immediately asked how she was doing. The following conversation ensued:

Steve: How are you?
Luanne: [*pause*] I'm feeling frightened and vulnerable, and like you're learning more about me than I realized I would be sharing today.
Steve: OK, I'm so glad you said that. So before we do any more, let's regroup. Is the level of feeling pretty intense right now?
Luanne: It's escalating, but that feels appropriate given the four pictures you showed me.
Steve: Exactly. I gave them in that order because I imagined that might happen.
Luanne: Yes, I thought so.
Steve: So, I know this is different than in therapy, but is what's happening now relevant to your question about which therapy approach is best for you?
Luanne: Hmmm . . . I'm not making the mental connection. I don't get the analogy.
Steve: OK . . . let me explain more. It's like we've got some different options here. We could keep going, with some harder cards, and intensify the feelings even more, or we could stop here, and call it a day, or perhaps talk some more about what we've already done. And that choice seems similar to me to your question about what therapy approach is best for you. I know it's a different situation, but it seems related to me. Does it to you?

At that point, Luanne confessed that she did not see it as a real option to stop. She just assumed that she had to go on, no matter how difficult it was. Stopping would feel like "weakness." Steve asked if stopping could be "a kindness to oneself" instead of weakness, and Luanne confessed that was a whole new way of thinking for her.

This interchange then led to a long discussion about the pros and cons of "pushing for feelings" versus "pacing oneself." Steve asked if there were unpleasant feelings that came up if Luanne gave herself breaks and didn't push so hard. She said, "There's sadness, and I'm more in touch with what I missed out on and what I long for." Steve asked her what she longed for, and Luanne paused before saying, "Love and support." Steve put his hand over his heart and made a sympathetic noise, and Luanne became tearful and looked away. Steve waited a moment and then said,

> Luanne, that makes so much sense. And yet, I have to tell you . . . when you were pushing ahead just now with the early cards, I had no idea how hard it was for you. It didn't occur to me to offer you any support. It was only when we stopped and you told me how vulnerable you felt that I had any inkling that you were in distress.

Luanne nodded and said that friends often told her that they could not tell when she was upset. Steve said,

> Well, I think it was unsafe growing up to show any weakness, and also you figured out you got more for yourself by just soldiering on. But now, in a way, when you do that, you miss out on what you most long for, because no one can even tell that you need support.

Luanne, sat quietly at that point for several minutes and said she had never thought about things that

way. Steve asked how she was feeling and she said, "Excited. Less frightened. I've got some options here that feel good to me."

At that point, Steve asked Luanne if she would like to look at more picture story cards or stop where they were. He warned her that the next cards were "even harder hitting emotionally" and might be difficult. Luanne said it was virtually impossible for her to imagine *not* going forward unless Steve had the sense that it would be bad for her and made the decision to stop. Steve felt the temptation to decide but declined, saying that he did not know what was best for Luanne at the moment. Instead, he asked if she wanted help "thinking through how to make such a decision." She said she definitely needed help, as she had no idea how to choose. Steve then shared several questions that he might ask himself in making such a decision, such as "How close am I to my emotional limit?" "What do I need to do after this—do I need to be 'on'?" and "How safe do I feel?" Luanne listened and answered each question for herself. She then paused and said, "I would like to do another card." Steve agreed and gave her a card showing a person kneeling, bent over on some kind of couch or cushion (TAT card 3BM). Luanne told a story of a woman who had been robbed and was terrified but who went to the police and got help. She started to blame herself, but the police reassured her she was not at fault. Eventually, she got the things back that had been stolen from her.

Steve asked Luanne what she noticed, and she said that the woman started out terrified but then ended up taking action and getting support and that made her feel better. Steve asked how Luanne was feeling, and she said "Good. I feel I learned a lot here today. And I've had enough now. I want to stop." Steve commended her for knowing that, and they ended.

Summary/Discussion Session

The summary/discussion session provides the opportunity for the client and assessor to collaboratively discuss the findings of the assessment. The assessor first contacts the referring professional to discuss the findings and plan the summary/discussion session together. Whenever possible, the referring therapist attends the summary/discussion session, and it is held at the therapist's office. The therapist sits with the client during the session, hears what the client hears, asks questions for the client, and "holds" the client emotionally during the process.

During the session, the assessor sets the client at ease as much as possible. Then the assessor takes each of the client's questions and proposes tentative answers based on the testing and previous discussions with the client. After each point, the assessor asks how the client understands the finding. If it seems to fit, the client is asked to provide examples of it in his or her life. The assessor stays attuned to any clue that the client is not following the discussion or that the test results do not seem accurate. The assessor is also attentive to the client going into shame or becoming overwhelmed. Often, the session ends with the client and therapist discussing viable next steps that the client can take to address the problems focused on in the assessment and talking about what it was like to do the assessment together.

Summary/Discussion Session: Case Illustration—Consultation Meeting With Sarah

Steve met with Sarah several days after the assessment intervention session with Luanne to bring her up to date, review the MMPI–2 results, answer Sarah's questions, and get her input on what he proposed to say to Luanne. Steve showed Sarah Luanne's MMPI–2 and shared his conceptualization of Sarah as a "sturdy survivor" who had to "shut off weakness" and "keep plunging ahead." Sarah, like Steve, was surprised at the level of distress Luanne revealed through the MMPI–2. Steve suggested that Luanne might shift her alliance more from Mary to Sarah if she felt that Sarah both recognized her distress and was attentive to Luanne's shame about appearing "weak." They both agreed that Sarah could avoid power struggles over the pacing of therapy by emphasizing that only Luanne could really know what was best for her in a given session but that Sarah could ask helpful questions and share her own impressions how to proceed. They agreed that Sarah's job was to help Luanne learn how to make

Exhibit 26.1
Excerpts From Written Feedback to Luanne at the End of the Assessment

Dear Luanne,

This is the letter I promised you, summarizing the results of the recent psychological assessment we did together. . . . I'll structure this letter as I did that session by addressing the questions you posed at the beginning of the assessment.

Before getting to the results, I want to thank you again for letting me get to know you through the assessment. I really enjoyed working with you and I appreciated the openness with which you approached the assessment. . . .

Now to your questions:

What still gets in my way of dating?

You told me that you have a lot of fear when you contemplate dating, and your MMPI–2 profile helps us understand that. Your test results showed that you often feel inferior, self-doubting, and are very self-critical. Although you are aware of your shame, and recognize it as a problem, at this point you still tend to believe it. This suggests that a part of you thinks you will be rejected by desirable men, which gets you to avoid dating or to pursue men who are unavailable—in a way taking control of the rejection.

There may even be some very practical reasons that dating is scary for you. Your MMPI–2 results suggest that you don't feel comfortable being assertive (more than the average woman psychotherapy client). People with scores like yours "lose their power" in relationships and have difficulty setting and maintaining appropriate boundaries. . . . So until you're able to be assertive and hold your own, you'll want to be cautious about whom you date. . . .

We also discussed your being a *sociable introvert* and how that complicates dating. You like and enjoy being around people, but are shy and can get anxious in social situations. You also need and cherish time alone. You might get along best with a partner who is similar to you in this regard (which is fairly rare) or at least you will need a partner who understands when you need time alone or are exhausted by lots of social contact.

Last we talked about how all the difficult feelings you are managing—of anxiety, depression, and shame—can make it difficult to date. We agreed that if you're able to date now, that's great. But if you can't, perhaps you can have compassion for yourself right now, and believe that as you feel better, dating could be a lot easier.

Is there a way for me to not be as controlled as I am by shame?

As mentioned earlier, the MMPI–2 does confirm that you feel bad about yourself and are very self-critical. In my experience, healing from shame involves the following steps: 1) being able to identify and label shame, especially when it is happening, and being curious about it; 2) learning about how your shame got there, what purposes it served, and developing compassion for yourself, especially so you don't feel shame about having shame; 3) letting yourself feel grief (including anger and sadness) about the circumstances that led you to feel shame, and the opportunities you've missed because of your shame; and 4) finding more and more skepticism for the self-critical thoughts and inner voices, so they no longer have much power over you.

From what you told me, it seems that you're currently working on step 2 and have a good handle on step 1. As we discussed, this entire process can take a long time, and in many ways is never done, so you'll need to be tender and patient with yourself.

What's the best approach for me in therapy, pushing for lots of feeling or a more paced, controlled approach? Is it possible and/or desirable to integrate these two approaches to treatment?

As we discussed and practiced in our second session (where we used the picture-story cards), there probably is no therapy approach that is best for you on every day and in every situation. Instead, the optimal pacing for your work will vary from day to day, depending on such factors as 1) your level of general emotional distress, 2) the demands of your life at the time, 3) the amount of support that is available to you, and 4) how much you want to push vs. give yourself a break. By giving yourself permission each day to decide how fast and deeply you want to delve, you'll best meet your goal of working as rapidly as you can, without getting disorganized or threatening your sobriety. And remember, this approach does NOT mean that you have to make these decisions all alone. . . .

In closing, thank you again, Luanne, for letting me get to know through the assessment, and I hope it is helpful to you in your future work with Sarah. If you have any questions about the assessment or about this letter, please feel free to call or email me, or to pass on the questions to me through Sarah.

Best wishes,
Steve
Stephen E. Finn, PhD
Licensed Psychologist

such decisions and to accept that Luanne might make some mistakes along the way.

The content of what Steve told Luanne is reflected in the letter he sent to her after the session was completed (see Exhibit 26.1). Here we describe the flow and process of the session.

Steve began the session by asking how Luanne felt after the assessment intervention session. Luanne said she was exhausted immediately afterward but since then had been feeling, "joyful," "more connected to herself," and that she "was going somewhere emotionally."

As can be seen in the letter to Luanne, Steve used Luanne's question about why she was not dating to talk about the large amount of distress shown on her MMPI–2, including the intense shame. He said that it was remarkable that Luanne was doing all she was doing and that this showed a lot of psychological strength. Luanne confirmed that she was struggling with very painful feelings and that she did not know what else to do besides "carry on." At that point, Sarah said that she thought Luanne often underestimated the amount of energy it was taking to keep going. Luanne responded that she felt like crying but did not want to. She then said she was glad that Sarah understood the effort she was making.

Steve then addressed Luanne's question about shame. Luanne listened intently as he described steps involved in healing from shame and said she did not have much experience "grieving what had happened to her." Steve asked if the good feelings Luanne had experienced since the previous session might be related to her having felt the sadness. Luanne said that might be true and that she felt more alive for having "come close to the edge and survived."

Next, Steve summarized the work he and Luanne had done about the best way to approach her therapy, reminding Luanne of the questions she might need to ask herself before and during each session. Sarah said she thought such questions were an excellent guide and that she could assume the role of helping Luanne "make her own decision." Steve got the sense that Luanne and Sarah were finally on the same page about the treatment. At the end of the session, Luanne said that the assessment had been a very "rich experience" and that she would be "feeding off it" for a very long time. Steve asked about what she felt she had learned, and Luanne said that "in complex situations, no one can decide what's best for me, but that doesn't mean I am all alone."

Written Feedback

After the summary/discussion session, the assessor writes a letter to the client that outlines the findings of the assessment that were discussed in the last session. Typically, it is in the form of a personal letter, which restates each question and summarizes the answer. This letter is an enduring documentation of the assessment findings and of the client's connection with the assessor.

Written Feedback: Case Illustration

The letter Steve sent to Luanne is excerpted in Exhibit 26.1.

Follow-Up Session

A follow-up session is typically scheduled 3 to 6 months after the summary/discussion session. It offers the opportunity for assessor and client to check in with each other and clarify or deepen what the assessment results indicate and how they might bear on recent questions and concerns. The follow-up serves as a mechanism to keep the client on track with the important results of the assessment. Sometimes the client requests additional follow-up sessions, and in some instances they become an annual occurrence.

Follow-Up Session: Case Illustration

Because of her busy school schedule, Luanne apologetically declined to come for a follow-up session. However, she returned a set of client feedback forms that Steve sent with the feedback letter and rated herself as highly satisfied with the assessment. When Steve checked with Sarah several months after the assessment, she told him that Luanne and she were working well together and that Luanne had recently been having almost no contact with her previous therapist, Mary. Sarah said she felt the assessment had helped her and Luanne bond in new and important ways.

CONCLUSIONS

We hope that this overview of TA and our case illustration convey the power and potential value of TA to psychological assessment and to clinical work in general. At this point, TA shows enormous promise to "breathe new life" into psychological assessment and to enhance our understanding of psychotherapy. Future research and clinical experience will determine whether TA lives up to this promise. Current research is focused on exploring TA's usefulness with different types of clients in diverse settings and in understanding why and for whom TA is most useful.

There is a vibrant assessment community that continues to explore TA's value, its applications, and ways to make it even more effective. Many members of this community are active in the Society for Personality Assessment (http://www.personality.org) and come together at its annual meeting. The TA website (http://www.therapeuticassessment.com) also contains more information about TA and lists of upcoming trainings.

References

Ackerman, S. J., Hilsenroth, M. J., Baity, M. R., & Blagys, M. D. (2000). Interaction of therapeutic process and alliance during psychological assessment. *Journal of Personality Assessment, 75*, 82–109. doi:10.1207/S15327752JPA7501_7

Ben-Porath, Y. S., Hostetler, K., Butcher, J. N., & Graham, J. R. (1989). New subscales for the MMPI–2 social introversion (Si) scale. *Psychological Assessment, 1*, 169–174. doi:10.1037/1040-3590.1.3.169

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (MMPI Restandardization Committee). (1989). *Manual for administration and scoring: MMPI–2*. Minneapolis: University of Minnesota Press.

Caldwell, A. B. (2001). What do the MMPI scales fundamentally measure? *Journal of Personality Assessment*, *76*, 1–17. doi:10.1207/S15327752JPA7601_1

Finn, S. E. (1996). *Manual for using the MMPI–2 as a therapeutic intervention*. Minneapolis: University of Minnesota Press.

Finn, S. E. (2007). *In our clients' shoes: Theory and techniques of therapeutic assessment*. Mahwah, NJ: Erlbaum.

Finn, S. E. (2009). The many faces of empathy in experiential, person-centered, collaborative assessment. *Journal of Personality Assessment, 91*, 20–23. doi:10.1080/00223890802483391

Finn, S. E., & Martin, H. (1997). Therapeutic assessment with the MMPI–2 in managed health care. In J. N. Butcher (Ed.), *Objective psychological assessment in managed health care: A practitioner's guide* (pp. 131–152). New York, NY: Oxford University Press.

Finn, S. E., & Tonsager, M. E. (1992). The therapeutic effects of providing MMPI–2 test feedback to college students awaiting psychotherapy. *Psychological Assessment, 4*, 278–287. doi:10.1037/1040-3590.4.3.278

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9*, 374–385. doi:10.1037/1040-3590.9.4.374

Fischer, C. T. (1978). Collaborative psychological assessment. In C. T. Fischer & S. L. Brodsky (Eds.), *Client participation in human services* (pp. 41–61). New Brunswick, NJ: Transaction Books.

Fischer, C. T. (1994). *Individualizing psychological assessment*. Mahwah, NJ: Erlbaum. (Original work published 1985)

Fosha, D. (2000). *The transforming power of affect: A model for accelerated change*. New York, NY: Basic Books.

Gorske, T. T., & Smith, S. (2008). *Collaborative therapeutic neuropsychological assessment*. New York, NY: Springer.

Graham, J. R. (2006). *MMPI–2: Assessing personality and psychopathology* (4th ed.). New York, NY: Oxford University Press.

Handler, L. (1999). The assessment of playfulness: Hermann Rorschach meets D. W. Winnicott. *Journal of Personality Assessment, 72*, 208–217. doi:10.1207/S15327752JP720205

Handler, L. (2006). Therapeutic assessment with children and adolescents In S. Smith & L. Handler (Eds.), *Clinical assessment of children and adolescents: A practitioners' guide* (pp. 53–72). Mahwah, NJ: Erlbaum.

Hilsenroth, M. J., Peters, E. J., & Ackerman, S. J. (2004). The development of therapeutic alliance during psychological assessment: Patient and therapist perspectives across treatment. *Journal of Personality Assessment, 83*, 332–344. doi:10.1207/s15327752jpa8303_14

Kohut, H. (1982). Introspection, empathy, and the semi-circle of mental health. *International Journal of Psychoanalysis, 63*, 395–407.

Kohut, H. (1984). *How does analysis cure?* Chicago, IL: University of Chicago Press.

Little, J. A., & Smith, S. R. (2008, March). *Collaborative assessment, supportive psychotherapy, or treatment as usual: An analysis of ultra-brief individualized intervention with psychiatric inpatients*. Paper

presented at the annual meeting of the Society for Personality Assessment, Chicago, IL.

Morey, L. C., Lowmaster, S. E., & Hopwood, C. J. (2010). A pilot study of manual-assisted cognitive therapy with a therapeutic assessment augmentation for borderline personality disorder. *Psychiatry Research, 178*, 531–535. doi:10.1016/j.psychres.2010.04.055

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Newman, M. L., & Greenway, P. (1997). Therapeutic effects of providing MMPI–2 test feedback to clients in a university counseling service: A collaborative approach. *Psychological Assessment, 9*, 122–131. doi:10.1037/1040-3590.9.2.122

Ougrin, D., Ng, A. V., & Low, J. (2008). Therapeutic assessment based on cognitive-analytic therapy for young people presenting with self-harm: Pilot study. *Psychiatric Bulletin*, *32*, 423–426. doi:10.1192/pb.bp.107.018473

Poston, J. M., & Hanson, W. M. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment, 22*, 203–212. doi:10.1037/a0018679

Purves, C. (2002). Collaborative assessment with involuntary populations: Foster children and their mothers. *Humanistic Psychologist, 30*, 164–174. doi:10.1080/08873267.2002.9977031

Schore, A. (2009). Right brain affect regulation: An essential mechanism of development, trauma, dissociation, and psychotherapy. In D. Fosha, M. Solomon, & D. Siegel (Eds.), *The healing power of emotions: Integrating relationships, body, and mind. A dialogue among scientists and clinicians* (pp. 112–144). New York, NY: Norton.

Silverton, L. (1993). *Adolescent Apperception Cards: Manual.* Los Angeles, CA: Western Psychological Services.

Smith, J. D., Handler, L., & Nash, M. R. (2010). Family therapeutic assessment for preadolescent boys with oppositional defiant disorder: A replicated single-case time-series design. *Psychological Assessment, 22*, 593–602. doi:10.1037/a0019697

Sotile, W. M., Henry, S. E., & Sotile, M. O. (1988). *Family Apperception Test: Manual.* Charlotte, NC: Feedback Services.

Swann, W. B., Jr. (1997). The trouble with change: Self-verification and allegiance to the self. *Psychological Science, 8*, 177–180. doi:10.1111/j.1467-9280.1997.tb00407.x

Swann, W. B., Jr., Wenzlaff, R. M., Krull, D. S., & Pelham, B. W. (1992). The allure of negative feedback: Self-verification strivings among depressed persons. *Journal of Abnormal Psychology, 101*, 293–306. doi:10.1037/0021-843X.101.2.293

Tharinger, D. J., Finn, S. E., Austin, C., Gentry, L., Bailey, E., Parton, V., & Fisher, M. (2008). Family sessions in psychological assessment with children: Goals, techniques, and clinical utility. *Journal of Personality Assessment, 90*, 547–558. doi:10.1080/00223890802388400

Tharinger, D. J. Finn, S. E., Gentry, L. B., & Matson, M. (in press). Therapeutic Assessment with adolescents and their parents: A comprehensive model. In D. Saklofske & V. Schwean (Eds.), *Oxford handbook of psychological assessment of children and adolescents.* New York, NY: Oxford University Press.

Tharinger, D. J., Finn, S. E., Gentry, L., Hamilton, A., Fowler, J., Matson, M., . . . Walkowiak, J. (2009). Therapeutic Assessment with children: A pilot study of treatment acceptability and outcome. *Journal of Personality Assessment*, *91*, 238–244. doi:10.1080/00223890902794275

Tharinger, D. J., Finn, S. E., Hersh, B., Wilkinson, A., Chistopher, G., & Tran, A. (2008). Assessment feedback with parents and children: A collaborative approach. *Professional Psychology: Research and Practice, 39*, 600–609. doi:10.1037/0735-7028.39.6.600

Tharinger, D. J., Finn, S. E., Wilkinson, A. D., DeHay, T., Parton, V., Bailey, E., & Tran, A. (2008). Providing psychological assessment feedback with children through individualized fables. *Professional Psychology: Research and Practice, 39*, 610–618. doi:10.1037/0735-7028.39.6.610

CHAPTER 27

ASSESSMENT OF GENDER-RELATED TRAITS, ATTITUDES, ROLES, NORMS, IDENTITY, AND EXPERIENCES

Bonnie Moradi and Mike C. Parent

Gender is socially constructed as a ubiquitous marker of human identity and experience (e.g., West & Zimmerman, 1987), and it intersects with other social categories (e.g., ability status, age, ethnicity, race, sexual orientation, social class) in ways that shape societal hierarchies of privilege and oppression (e.g., Collins, 1990; Johnson, 2006; West & Fenstermaker, 1995). Despite the socially construed significance of gender, the assessment of gender-related constructs has evolved with notable shifts and variability in fundamental theoretical assumptions. Our vision for this chapter is to highlight selected classes of gender-related constructs with the goal of promoting conceptual clarity in the meaning and operationalization of these constructs. We hope to achieve this goal by (a) offering an organizational scheme of classes of gender-related constructs and corresponding measures and, (b) within each class, describing selected measures that exemplify that class of constructs. The order of presentation of gender-related constructs and measures in this chapter roughly reflects their conceptual evolution: from unilinear to bilinear to multidimensional, and from trait- or personality-related individual difference dimensions to attitudes toward traditional gender norms to collective consciousness and experiences.

We begin with a critical discussion of some underlying assumptions in psychological conceptualizations of sex, gender, and related constructs, as these assumptions are implicit in most current approaches to operationalizing gender-related constructs. We then present seven classes of gender-related constructs, describing example measures of each class. We conclude with a call for construct clarity, measurement consolidation and refinement, and integration of transgender issues and intersecting identities in the assessment of gender-related constructs.

SEX AND GENDER

Feminist scholars have distinguished *sex*, or the biological aspects of female and male sex categories, from *gender*, or the socially constructed meanings (e.g., traits, roles, behaviors) afforded to these sex categories (American Psychological Association [APA] Task Force on Gender Identity and Gender Variance, 2008; Bem, 1993; West & Zimmerman, 1987). However, in the psychological literature, sex and gender descriptors are sometimes used interchangeably without clarity or consistency about their conceptual meaning or operationalization. A common manifestation of this confusion is the use of sex category descriptors (i.e., female and male) when actual biological markers of sex are not assessed; rather, participants' self-reports of their gender group identification, which may or may not correspond with their biological sex characteristics, are assumed to be accurate proxies of biological sex.

The imperfect correspondence of biological sex characteristics and gender group identification is underscored by the fact that the biological characteristics that underlie sex categories are not dimorphic; rather, they are multidimensional and multicategorical (perhaps even continuous). For example, sex chromosomes; the presence and

DOI: 10.1037/14048-027
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology, K. F. Geisinger (Editor-in-Chief)

functionality of gonads; the relative balance of gonadal hormones; and the presence, size, and functionality of internal and external sex organs each exist with much greater variability (both within and between individuals) than is suggested by the practice of dichotomous sex category assignment typically based on phallus size (for a reader-friendly overview, see Yoder, 2007; for a more in-depth review of sex and gender diversity across species, see Roughgarden, 2004). Also, people may adopt a gender identity that differs from their assigned sex category or may eschew gender category identification altogether, with or without changing their sex characteristics (for a review, see APA Task Force on Gender Identity and Gender Variance, 2008). In light of these complexities, greater attention is needed to intersex and transgender issues and multiple sex and gender categories in conceptualization, assessment, and reporting of sex and gender (e.g. APA Task Force on Gender Identity and Gender Variance, 2008; Fausto-Sterling, 1993).

Despite these challenges to the use of dichotomous and mutually exclusive sex and gender categories (i.e., female/woman, male/man), many of the current approaches to assessing gender-related constructs reflect an implicit dimorphic view of sex and gender and assume correspondence between the two. Thus, in reviewing each of the classes of measures discussed in the proceeding sections, it is helpful to consider the extent to which implicit dimorphism shapes and is reinforced by the conceptual underpinnings of the measures. An additional context for the proceeding discussion is that social construction of gender does not occur in isolation; rather, gender is coconstructed with other social categories. Thus, in reviewing the classes of measures, it is important to consider the extent to which the measures capture potential variability in gender-related constructs across age, ability status, ethnicity, race, sexual orientation, and other social categories or the extent to which the measures justify a particular frame of reference in their conceptual underpinnings. Both of these considerations—implicit dimorphism and intersections of gender with other dimensions of identity—are an important context for the proceeding review of classes of gender-related measures.

CLASSES OF GENDER-RELATED CONSTRUCTS AND REVIEW OF SELECTED MEASURES

Gendered Personality Dimensions

Early measures operationalized gender as a unidimensional trait with masculinity and femininity as opposing poles of a single continuum. Such measures were critiqued for precluding the coexistence of masculinity and femininity and reflecting an essentialist construal of gender as a dispositional trait rather than a social construction (e.g., Morawski, 1985; Smiler, 2004). Although these measures have been eclipsed by modern multidimensional conceptualizations of gender, research with this class of measures constitutes a large part of the gender literature and the measures are still employed in research programs and personality assessment batteries.

The most widely administered unilinear trait measure of gender may be Scale 5, Masculinity-Femininity (Mf) of the Minnesota Multiphasic Personality Inventory and its major revision (MMPI and MMPI–2, respectively; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The Mf scale is one of the standard 10 MMPI–2 scales and is obtained with all standard scoring protocols (Wong, 1984). The Mf scale was originally developed to detect homosexuality on the basis of an implicit hypothesis of gender inversion among gay men and lesbians (Hathaway & McKinley, 1940; Singer, 1970); subsequent research with the Mf scale indicated that it performed this task very poorly (for a review, see Wong, 1984). Validity investigations indicate that, at best, the Mf scale assesses stereotypical gendered interests broadly (perhaps better so for women than for men) and has limited utility in clinical settings (e.g., Lewin & Wild, 1991; Wong, 1984).

In part as a critique of unilinear trait measures, the Bem Sex Role Inventory (BSRI; Bem, 1974) and the Personal Attributes Questionnaire (PAQ; Spence & Helmreich, 1978; Spence, Helmreich, & Stapp, 1973) were developed to operationalize masculinity and femininity as related but distinct continua, allowing for individuals to score high or low on both dimensions. This shift from unilinear to bilinear assessment was an important advancement in theory

and research on gender. The BSRI and PAQ reflected somewhat different theoretical aims—to operationalize gender schema theory and gender-related personality traits, respectively. For both measures, however, items were selected based on college students' ratings of characteristics that were desirable (and typical in the case of the PAQ) for women and men (for a review, see Spence, 1991). The BSRI Masculine and Feminine subscales each comprise 20 items that can be assessed continuously or used to classify individuals into masculine-typed, feminine-typed, androgynous (high on masculine and feminine), or undifferentiated (low on masculine and feminine; Bem, 1977) categories. In the typically used version of the PAQ, eight items each assess masculinity, femininity, and androgyny, although the androgyny items are frequently not scored (Spence, 1991; Spence & Helmreich, 1978).

Although the BSRI and PAQ reflected important advancements in the assessment of gender-related constructs, both measures have been critiqued on several grounds. For example, the theoretical bases for item selection, evaluation, and retention in these measures have been questioned (e.g., Spence, 1991). A critique of the BRSI also has been that its masculinity items are more generally desirable than its femininity items (e.g., Hoffman & Borders, 2001; Pedhazur & Tetenbaum, 1979). Furthermore, investigations of the factor structure of these measures have raised concerns about the replicability of a two-factor structure and suggest that instead of assessing masculinity and femininity, the BSRI and PAQ assess emotional expressiveness and instrumentality/autonomy, other personality-like characteristics, or self-esteem (e.g., Choi & Fuqua, 2003; Nicholls, Licht, & Pearl, 1982; Ward, Thorn, Clements, Dixon, & Sanford, 2006). Finally, as with the Mf scale, a critique of the BRSI and PAQ is that, by operationalizing gender as traits, they reflect an essentialist rather than a social–constructionist view of gender (Smiler, 2004).

Areas for advancement. Early trait measures were critical in the evolution of assessing gender-related constructs. However, the unilinear and essentialist roots of the Mf scale and ample critiques of its validity render it inappropriate for assessing gender or the relation of gender with psychological symptoms. Similarly, use of the Mf scale to diagnose homosexuality is empirically unsupported, is pathologizing, and it contradicts the standard of affirmative approaches to sexual minority groups and identities (e.g., APA Division 44/Committee on Lesbian, Gay, and Bisexual Concerns Joint Task Force, 2000; APA Task Force on Appropriate Therapeutic Responses to Sexual Orientation, 2009; Fox, 1988; Wong, 1984). Although the BSRI and PAQ have the strength of assessing gender bilinearly, critiques of these measures also suggest their cautious and circumscribed use as brief measures of instrumentality and emotional expressiveness. In assessing gender-related constructs, other measures of the precise construct of interest, such as those reviewed in the proceeding sections, may be more appropriate.

Attitudes Toward Women, Men, and Transgender Individuals

One of the largest classes of gender-related measures is those that assess attitudes toward women and men, often relative to one another. Some of the measures in this class focus on assessing attitudes regarding equal rights, roles, and responsibilities, whereas others focus on assessing prejudicial attitudes. Measures of attitudes toward transgender people are also emerging and broaden the scope of the literature beyond dimorphic views of gender.

Attitudes toward rights and roles of women and men. The Attitudes Toward Women Scale (AWS; Spence & Helmreich, 1972) is one of the most widely used measures of gender-related constructs (Spence & Hahn, 1997); its most frequently used version is the 15-item form (Spence & Helmreich, 1978). Long-standing use of the AWS provides ample psychometric data regarding this measure. There is support for the stability of a unidimensional structure and Cronbach's alphas in the 0.80s over the past several decades (e.g., Spence & Hahn, 1997; Whatley, 2008). There are also plenty of validity-consistent findings, as the AWS is typically used as a convergent validity indicator for new measures of gender-related constructs (reviewed in subsequent sections). There is also evidence that AWS scores have been rising

over time (e.g., Spence & Hahn, 1997; Twenge, 1997), suggesting an increase in egalitarian attitudes toward women. This evidence, however, magnifies concerns about ceiling effects that limit the utility of the AWS for differentiating attitudes in the liberal end of its continuum and for assessing modern and subtle sexist attitudes (e.g., Beere, 1990; Fassinger, 1994; Spence & Hahn, 1997).

Several measures have been developed to address concerns about the AWS and to assess more nuanced constructs within the domain of attitudes toward the rights and roles of women and men. A conceptually close advancement to the AWS is the Sex-Role Egalitarianism Scale (SRES; Beere, King, Beere, & King, 1984) designed to assess attitudes toward the equality of women and men in five domains: marital, parental, employment, social/relational, and educational roles. SRES items were developed on the basis of conceptual grounding and then submitted to psychometric evaluation with data from diverse populations, including police officers, senior citizens, and students. These analyses resulted in four SRES forms: two alternate 95-item versions and two alternate 25-item versions (for a review, see King & King, 1997).

Cronbach's alpha and 2- to 3-week test–retest reliability coefficients were in the 0.80s and 0.90s for SRES overall scale and subscales across all forms (King & King, 1997). Subsequent studies have yielded comparable Cronbach's alphas with such diverse samples as police officers, university faculty and staff, and managers and executives and in studies with Italian and Croatian translations of the long form of the SRES (King & King, 1997). King, King, Gudanowski, and Taft's (1997) confirmatory factor analyses of SRES data for both 95-item versions supported the five-factor structure corresponding with the originally posited five domains. These analyses also supported a hierarchical structure, with Intimate Relationship Egalitarianism as a higher order factor over marital, parental, and social/relational roles and Formal Relationship Egalitarianism as a higher order factor over employment and educational roles. Item-response theory-based analyses of SRES data have suggested that the SRES is most effective in distinguishing low from neutral or high egalitarian attitudes (Vreven, King, & King, 1994).

Since its development, the SRES has been used in a wide range of populations and research topics, yielding validity-consistent findings. For example, SRES scores indicated more egalitarian attitudes among feminist than nonfeminist women, among women than men, and among younger individuals than older individuals; SRES scores were independent from socially desirable responding and BSRI scores, were linked positively with AWS scores, and were linked negatively with indicators of marital discord and violence (King & King, 1997). Overall, the SRES is one of the more extensively evaluated scales in its class of measures and it remains a useful tool for assessing attitudes toward the relative rights and roles of women and men.

Ambivalent attitudes toward women and men. Another approach to assessing attitudes toward women and men is represented in the Ambivalent Sexism Inventory (ASI; Glick & Fiske, 1996) and the Ambivalence Toward Men Inventory (AMI; Glick & Fiske, 1999). Glick and Fiske (1996, 1999) theorized that the interpersonal dynamics between women and men in a patriarchal society can result in a mix of hostile and benevolent attitudes. Hostile sexism toward women comprises dominative paternalism, competitive gender differentiation, and heterosexual hostility, whereas benevolent sexism toward women comprises protective paternalism, complementary gender differentiation, and heterosexual intimacy. Hostile and benevolent sexism are thought to work together to legitimize men's power and status over women and to prescribe and reinforce subordinate and nonthreatening roles for women. This patriarchal power structure can, in turn, promote hostile and benevolent attitudes toward men along parallel dimensions. Hostile attitudes toward men reflect resentment of paternalism, compensatory gender differentiation, and heterosexual hostility, whereas benevolent attitudes toward men reflect maternalism, complementary gender differentiation, and heterosexual attraction.

Item pools were developed to capture this conceptual content and then reduced to 22 ASI items and 20 AMI items based on principal-components analyses of data from college women and men (Glick & Fiske, 1996, 1999). Confirmatory factor

analyses also were conducted with ASI and AMI data from multiple samples. For ASI data, these analyses suggested two higher order factors, Hostile Sexism and Benevolent Sexism, and three second order Benevolent Sexism factors, Protective Paternalism, Complementary Gender Differentiation, and Heterosexual Intimacy; this structure provided better fit than a one- or two-factor model but was not supported uniformly across samples (Glick & Fiske, 1996). Similarly, confirmatory factor analyses of AMI data suggested two higher order factors of hostile and benevolent attitudes, each with three second order factors reflecting the previously described components of these attitudes; this structure provided better fit than a one- or two-factor model but was not supported uniformly across samples (Glick & Fiske, 1999). Confirmatory factor analyses of cross-national data generally supported the hypothesized hierarchical model of the AMI; support was less consistent for the ASI, with a unidimensional model, two-factor model, or hierarchical model emerging as the best fitting model depending on the country (Glick et al., 2000, 2004). Across the scale development samples, Cronbach's alphas for ASI and AMI scale and subscale items ranged between 0.70s and 0.90s (Glick & Fiske, 1996, 1999); generally comparable alphas emerged across countries, but some subscale alphas dipped below 0.70 (Glick et al., 2000, 2004).

The initial instrument development studies and subsequent use of the ASI and AMI have produced substantial evidence of construct validity for these measures. For example, ASI scores are correlated positively with modern sexism and AWS scores but correlated weakly with impression management, and AMI scores are correlated positively with stereotypic attitudes toward men (Glick & Fiske, 1996; 1999). In cross-national data, ASI and AMI scores generally were correlated positively with each other and with gender stereotypes, and were associated negatively with countries' United Nations indices of gender equality (Glick et al., 2000, 2004). These data are consistent with the supposition that hostile and benevolent attitudes toward men and toward women are associated with each other and with societal disempowerment of women.

Attitudes toward transgender people and issues. The measures of attitudes toward women and men discussed so far fit with a dimorphic view of sex and gender. However, some recent efforts advance assessment of attitudes toward gender groups beyond the man–woman or male–female duality. For example, Hill and Willoughby (2005) developed the Genderism and Transphobia Scale (GTS). GTS items were developed on the basis of a review of the literature and then submitted to psychometric evaluation with data from Canadian college students. Items were selected on the basis of their correlations with subscale totals, and, ultimately, 32 items and three subscales were retained: Genderism, or acceptance of a sex and gender dichotomy; Transphobia, or disdain for gender transgression; and Gender-Bashing, or proclivity toward violence against gender transgressors. Cronbach's alphas for scale and subscale items ranged from 0.79 to 0.96 across three college and community samples. Principal-components analysis of data from the third sample suggested two factors: Gender-Bashing and Genderism/Transphobia. In terms of validity, GTS scores were correlated positively with homophobia and traditional gender role ideology. Furthermore, as expected, individuals who had met a transgender individual scored significantly lower on the GTS than did individuals who had not. GTS scores were not correlated with BSRI masculinity and femininity, self-esteem, or social desirability.

Nagoshi et al. (2008) noted some problems with the development of the GTS; mainly, that factor analysis did not inform instrument development and that high GTS subscale intercorrelations (*r*s over .70) challenged discriminant validity. To address these limitations, Nagoshi et al. (2008) adapted items from Bornstein's (1998) Flexibility of Gender Aptitudes and developed the nine-item Transphobia Scale (TPS). In their sample of college students, TPS items had a Cronbach's alpha of 0.82, and the hypothesized unidimensional structure was supported by a principal-components analysis. TPS scores correlated positively with hostile and benevolent sexist attitudes, right-wing authoritarianism, and religious fundamentalism and were unrelated to PAQ masculinity and femininity (Nagoshi et al., 2008).

Areas for advancement. The study of attitudes toward women and men has been fruitful, but the multitude of measures in this body of research suggests the need for conceptual clarity in measurement selection. For example, continued use of the AWS as a proxy for gender role ideology, feminist attitudes, and other constructs is problematic in light of the availability of more precise measures of such constructs and the concerns about ceiling effects in AWS scores. Although the SRES may share the limitation of discriminating attitudes at the egalitarian end of the continuum, it has the advantages of extensive psychometric evaluation, domain-specific assessment of egalitarian attitudes, and assessing attitudes toward both women and men. The ASI and AMI also reflect strong theoretical grounding and extensive psychometric evaluation. One critique of this line of research, however, is that little actual ambivalence may be captured in hostile and benevolent attitudes (Petrocelli, 2002); ambivalence suggests opposing beliefs, but hostile and benevolent attitudes are correlated positively with each other and generally correlated in the same direction with other variables. Nevertheless, the theoretical underpinnings of the ASI and AMI may provide a framework for organizing and consolidating other measures of sexism and attitudes toward women and men into higher order factors of hostile or benevolent attitudes. As well, further investigation of attitudes toward transgender individuals is needed to build on the burgeoning research in this area. Refining measures of transphobia can facilitate research on the implications of gender transgression for transgender and nontransgender people.

Gender Role Ideology

The aforementioned measures of attitudes toward women and men are sometimes used as proxies for gender role ideology, and they tap this construct to some extent. However, several measures are designed specifically to assess gender role ideology as beliefs that women and men should behave in particular, socially prescribed ways as a function of their gender.

The Male Role Norms Scale (MRNS; Thompson & Pleck, 1986) was developed using items from an older measure (Brannon & Juni, 1984). Data from college men were submitted to principal-components analysis on the basis of which the 26-item MRNS was formed to assess status, toughness, and antifemininity norms (Thompson & Pleck, 1986). Subsequent confirmatory factor analyses of data from U.S. college men indicated the presence of four factors: Tough Image, Violent Toughness, Status/Rationality, and Antifemininity (Fischer, Tokar, Good, & Snell, 1998). This four-factor model provided slightly better fit than the originally posited three-factor model in a sample of Turkish women and men students (Lease, Çiftçi, Demir, & Boyraz, 2009). Subsequent studies have linked MRNS scores with such variables as fear of emotions, negative attitudes toward gay men, and belief that rape accusations and laws are unfair to men (e.g., Holz & DiLalla, 2007; Jakupcak, Tull, & Roemer, 2005; Parrott, Peterson, Vincent, & Bakeman, 2008). Cronbach's alphas for MRNS subscale items have been over 0.70 in the aforementioned studies. Overall, the MRNS has yielded some validity-consistent evidence and acceptable reliabilities. However, factor analyses raise questions about its scoring along three subscales, and these subscales do not cover the range of norms identified in other masculine norms measures.

Another measure of masculine ideology is the Male Role Norms Inventory (MRNI; Levant & Fischer, 1998; Levant et al., 1992). The MRNI is a 57-item measure of endorsement of masculine ideology along seven domains: Avoidance of Femininity, Rejection of Homosexuals, Self-Reliance, Aggression, Achievement/Status, Attitudes Toward Sex, and Restrictive Emotionality. The MRNI also yields an overall Traditional masculine ideology score and an overall Nontraditional Attitudes score. Despite the posited multidimensionality of the MRNI, however, studies typically used the overall traditional or nontraditional masculine ideology scores rather than the specific subscale scores. Traditional MRNI scores were linked with such variables as relationship dissatisfaction, reluctance to talk about condoms with an intimate partner, and negative attitudes toward diversity and women's equality (for a review, see Levant & Richmond, 2007). However, psychometric concerns about the MRNI (e.g., factor structure, reliability) led to the development

of the MRNI–Revised (MRNI-R; Levant, Rankin, Williams, Hasan, & Smalley, 2010; Levant, Smalley, et al., 2007).

The MRNI-R consists of 53 reworded or new items that assess masculine ideology along seven dimensions, slightly modified relative to the original (Levant et al., 2007), and refined again based on principal axis factor analysis of data from college women and men (Levant et al., 2010); these seven factors are: Restrictive Emotionality, Self-Reliance Through Mechanical Skills, Negativity Toward Sexual Minorities, Avoidance of Femininity, Importance of Sex, Toughness, and Dominance. In Levant et al.'s (2010) samples of college women and men, Cronbach's alpha was 0.96 for all items and ranged from 0.75 to 0.92 for MRNI-R subscale items, indicating improvements over the original MRNI. In this sample, men scored significantly higher than women on all subscales, and MRNI-R total scores were correlated in expected directions with scores on measures of masculine gender role conflict and conformity to masculine norms; notably, however, MRNI-R scores were not correlated with the masculinity subscale of the PAQ. In other studies, MRNI-R total scores were correlated positively with family conflict, and correlated negatively with family cohesiveness, time spent with children, and attitudes toward seeking psychological help (Berger, Levant, McMillan, Kelleher, & Sellers, 2005; Boyraz & Sayger, 2009). Confirmatory factor analyses to test the structural stability of the MRNI-R are still needed. Psychometric investigations might also evaluate whether some subscales have too few items to cover the intended content domain (e.g., Importance of Sex contains only three items). Evaluating and preserving the intended multidimensionality of the MRNI-R is important as use of total scale scores obfuscates the theorized multidimensionality of masculine norms.

Relative to research on masculine gender role ideology, limited research has attended to feminine gender role ideology. One available measure of feminine gender role ideology is the Female Role Norms Scale (FRNS; Lefkowitz, Shearer, Gillen, & Espinosa-Hernandez, 2006). The seven-item FRNS is based on rewordings of MRNS antifemininity items to reflect disdain for women engaging in stereotypically masculine activities. FRNS scores have been associated with antigay and antilesbian attitudes, right-wing authoritarianism, social dominance orientation, and traditional attitudes toward marital roles in student samples of women and men; FRNS items yielded Cronbach's alphas of 0.76 to 0.79 in these samples (Gillen & Lefkowitz, 2006; Goodman & Moradi, 2008). However, psychometric evaluation and use of the FRNS remains limited.

As another measure of feminine gender role ideology, Philpot (as cited in Levant, Richmond, Cook, House, & Aupont, 2007) developed the Feminine Ideology Scale (FIS); although the FIS was used in research, its development was not published in a peer-reviewed form. Levant, Richmond, et al.'s (2007) description of the FIS indicates that its items were developed based on conceptual grounds and then submitted to psychometric evaluation using data from college women and men. Although some of the psychometric procedures seem problematic (e.g., a priori specification of five factors in a principal-components analysis with orthogonal rotation), they resulted in retention of 45 items that assess beliefs along five domains of feminine ideology: Stereotypic Images and Activities, Dependency/Deference, Purity, Caretaking, and Emotionality. Levant, Richmond, et al. (2007) undertook further psychometric analyses of the FIS with a sample of college women and men. In this investigation, FIS subscale items yielded Cronbach's alphas of 0.72 to 0.93, and principal-components analysis yielded results consistent with the expected five-factor structure. As evidence of validity, women and men's FIS subscale scores were correlated positively with MRNI scores and were generally independent of BSRI masculinity and femininity scores; women's FIS scores also were correlated positively with passive acceptance of sexism and traditional gender roles (Levant, Richmond, et al., 2007).

Areas for advancement. Research on gender role ideology and its assessment has been growing. However, feminine gender role ideology has received relatively less attention than has masculine gender role ideology. As such, further research on the FRNS and FIS, including examination of the factor structure of these measures and their applicability

to women as well as men is needed. Confirmatory factor analyses are also needed to address limitations in previous research on the structural properties of gender role ideology measures. Specifically, despite its noted utility in instrument development (e.g., Worthington & Whittaker, 2006), principal axis factor analysis appears to be underutilized in the gender role ideology literature, and principal-components analysis is used, often without clear justification. Similarly, despite conceptual and empirical grounds for correlated factors, orthogonal rotations are used frequently without description of potential oblique solutions. These concerns are also applicable to some of the other classes of measures reviewed, but they are noted here because confirmatory factor analyses of gender role ideology measures are sparse. Confirmatory factor analyses could evaluate the stability of previously observed structures and test the factorial invariance of the measures across gender groups, given that measures of gender role ideology are intended for use across groups.

Gender Role-Related Stress and Conflict

One of the most researched areas of gender-related constructs is the gender role stress and conflict paradigm, particularly as applied to masculine gender roles. This class of measures grew from the view that gender roles can limit women and men's potential and these measures aim to assess the conflict or stress that can arise from adopting restrictive gender roles (O'Neil, 2008).

Masculine gender role conflict is conceptualized as the incongruence between optimal functioning and conformity to social standards of behavior for men; the Gender Role Conflict Scale (GRCS; O'Neil, Helms, Gable, David, & Wrightsman, 1986) was developed to assess this conflict. The GRCS originally contained two versions: Personal Self-Report, assessing how men think and feel about gendered behaviors, and Situational Self Report, assessing potential conflict in situations in which gender is salient. The 37-item Personal Self-Report is the typically used version and it has four subscales: Success, Power, and Competition; Restrictive Emotionality; Restrictive Affective Behavior Toward Men; and Conflicts Between Work and Family Relations. The GRCS has been used with samples of men across North America, Europe, Australia, and Asia, and with participants diverse in age, race/ethnicity, and sexual orientation (for a review, see O'Neil, 2008). Cronbach's alphas of the GRCS subscale items have been in the 0.70s and 0.80s across samples; exploratory and confirmatory factor analyses have supported the hypothesized four factor structure; and validity evidence includes positive correlations with masculine gender role stress, masculine ideology, conformity to masculine norms, and psychological symptomatology (for a review, see O'Neil, 2008). Within the context of these psychometric strengths, however, a point of conceptual ambiguity is whether the GRCS assesses gender role conflict or self-attributed gender norms (Betz & Fitzgerald, 1993). For example, an item that assesses dislike of showing emotions to others reflects conflict less directly than does an item that assesses feeling torn between work and self-care. Thus, although items may assess constructs that are stressful for men, it is not clear whether all items specifically assess felt conflict over gender norms or if it is assumed that men who endorse gender norm items will be more likely to experience conflict.

The Masculine Gender Role Stress Scale (MGRS) assesses "cognitive appraisal of specific situations as stressful for men" (Eisler & Skidmore, 1987, p. 125). It is based on the premise that highly gender-conforming men will report greater anticipated stress for engaging in activities that are stereotypically feminine. College women and men were included in the initial item generation and selection procedures for the MGRS, but only men were included in the principal-components analysis that informed subscale formation. The 40-item MGRS includes five subscales: Physical Inadequacy, Emotional Inexpressiveness, Subordination to Women, Intellectual Inferiority, and Performance Failure. This hypothesized five-factor factor structure was supported in a confirmatory factor analysis, although this investigation was limited by the sample size of only 108 men (McCreary, Newcomb, & Sadava, 1999). Eisler and Skidmore (1987) found that men scored higher than women on overall MGRS scores. MGRS scores have been associated with such variables as negative attitudes toward gay men, drive for muscularity, and propensity for aggression and violence in intimate relationships

(Mahalik, Aldarondo, Gilbert-Gokhale, & Shore, 2005; Moore & Stuart, 2005; Mussap, 2008; Parrott et al., 2008).

Parallel to the MGRS, the Feminine Gender Role Stress Scale (FGRS; Gillespie & Eisler, 1992) assesses gender role socialization-related stressors for women. Generation and selection of items was based on data from college women and men, but only data from women were included in the principal axis factor analysis that informed subscale formation. The 39-item FGRS contains five subscales: Fear of Unemotional Relationships, Fear of Being Unattractive, Fear of Victimization, Fear of Behaving Assertively, and Fear of Not Being Nurturant. Cronbach's alphas have been acceptable (e.g., 0.73 and higher in the developmental study and in Mussap, 2008). Women's overall FGRS scores were correlated positively with depression, daily hassles, and PAQ femininity, but these correlations were not consistently significant at the subscale level (Gillespie & Eisler, 1992).

Both the MGRS and the FGRS have been translated and used with Chinese and Dutch women and men (Tang & Lau, 1995, 1996; van Well, Kolk, & Arrindell, 2005). In these samples, confirmatory factor analyses did not indicate acceptable fit for the hypothesized five-factor (Tang & Lau, 1996; van Well et al., 2005). On the basis of principal axis factor analysis of data from Chinese women and men, Tang and Lau (1996) suggested three factors for the MGRS (i.e., Performance Failure, Inferiority, and Emotional Inexpressiveness) and the FGRS (i.e., Inadequacy, Unassertiveness, and Victimization), but these models did not fit acceptably in the Dutch sample (van Well et al., 2005). In the Chinese and Dutch samples, Cronbach's alphas approximated the low 0.70s and 0.80s for MGRS and FGRS subscale items and the low 0.90s for scale items. MGRS and FGRS scores were associated with symptomatology and daily stress, but gender differences were more consistent on FGRS than on MGRS subscales, raising questions about the gender specificity of the stress assessed in the MGRS (Tang & Lau, 1995, 1996; van Well et al., 2005).

Areas for advancement. Research on gender role conflict and stress has produced useful findings for theory, research, and practice in a wide range of domains. However, a number of important areas for further investigation remain. Specifically, in light of some mixed support for differences between women's and men's scores (Tang & Lau, 1996; van Well et al., 2005), research is needed to evaluate the degree to which this class of measures assesses gender-specific conflict or stress. Relatedly, college men's GRCS and MGRS scores were found to overlap with Big Five personality dimensions and these personality dimensions, most typically neuroticism, mediated the links of GRCS and MGRS subscale scores with a number of mental health criterion variables (Tokar, Fischer, Schaub, & Moradi, 2000). As well, there is some evidence of overlap between this class of measures and BSRI and PAQ instrumentality and expressiveness scores (e.g., Gillespie & Eisler, 1992; van Well et al., 2005). These findings raise questions about the distinctiveness of gender role stress and conflict scores from personality dimensions. Similarly, evaluation of item content for this class of measures suggests the need to clarify their distinction from measures of gender role ideology and self-attributed gender norms (reviewed next).

Self-Attributed Gender Norms

Measures of self-attributed gender norms differ from gender role ideology measures in that they assess personal conformity to gender norms rather than ideology about those norms in general. The personal conformity measures also do not assume that conformity to gender norms is necessarily maladaptive or stress inducing. Specifically, Mahalik et al. (2003) argued that gender norms are influenced by dominant cultural values (e.g., White, heterosexual) to which all groups are held. Two measures were developed to assess respondents' conformity to the gender norms reflected in dominant U.S. cultural values: the Conformity to Masculine Norms Inventory (CMNI; Mahalik et al., 2003) and the Conformity to Feminine Norms Inventory (CFNI; Mahalik, Morray, et al., 2005), both of which have been abbreviated (Parent & Moradi, 2009, 2010).

The CMNI and CFNI were developed using rational instrument development procedures followed by psychometric evaluation and refinement. Mahalik et al. (2003; Mahalik, Morray, et al., 2005)

used focus groups of women and men to identify dominant U.S. cultural norms for how women and men should think, act, and feel; literature reviews were used to identify additional themes. Resultant item pools were administered to college students (women for the CFNI and men for the CMNI), and these data were used to conduct principal axis factor analyses. In each case, multiple dimensions were retained. The original 94-item CMNI comprised Emotional Control, Disdain for Homosexuals, Dominance, Power Over Women, Playboy, Primacy of Work, Pursuit of Status, Risk-Taking, Self-Reliance, Violence, and Winning. The original 84-item CFNI comprised Care for Children, Domestic, Sexual Fidelity, Invest in Appearance, Nice in Relationships, Modesty, Romantic Relationship, and Thinness. Men's CMNI scores have been linked with rape myth acceptance, negative attitudes toward help seeking, relationship dissatisfaction, propensity to cope with depression through drinking, and adaptive behaviors such as exercising to cope with depression (Burn & Ward, 2005; Locke & Mahalik, 2005; Mahalik, Burns, & Syzdak, 2007; Mahalik & Rochlen, 2006); women's CFNI scores have been related to greater body concerns, eating disorder symptoms, depression, and lower self-esteem (Hurt et al., 2007).

Parent and Moradi (2009, 2010) conducted confirmatory factor analyses of both the CMNI and CFNI with the aim of clarifying factor structure and abbreviating both measures. On the basis of their findings, Parent and Moradi suggested deletion of the CMNI Dominance and Pursuit of Status subscales, which had demonstrated relatively weak factor specificity, validity evidence, and reliability coefficients in previous research. As well, they suggested renaming the Disdain for Homosexuals subscale as Heterosexual Self-Presentation to reflect its item content more accurately. In the case of the CFNI, Parent and Moradi suggested separating items originally designed to load on to two factors (Sweet and Nice, Relational) that had been merged into a single Nice in Relationships factor during scale development. In each case, the abbreviated measures, CMNI-46 and CFNI-45, were approximately half the length of the originals and demonstrated superior data-model fit while retaining Cronbach's alphas comparable with the original form (0.77 to 0.91 for CMNI-46 subscale items; 0.68 to 0.89 for CFNI-45 subscale items). In subsequent studies, the factor structure of both abbreviated measures was replicated; convergent validity evidence was garnered in CMNI-46 and CFNI-45 subscale scores' positive correlations with scores on measures of parallel constructs; and discriminant validity evidence was garnered in CMNI-46 and CFNI-45 subscale scores' generally small correlations with such constructs as social desirability, Big Five personality dimensions, instrumentality, expressiveness, and self-esteem (Parent & Moradi, 2009, 2010, 2011a, 2011b; Parent, Moradi, Rummell, & Tokar, 2011).

Areas for advancement. Research on personal gender norm conformity has generated promising findings, but several important considerations remain. Specifically, much of the research with the CMNI and CFNI has used overall scale, rather than subscale, scores despite low to moderate subscale intercorrelations and factor-analytic evidence of the multidimensionality of data produced by these measures (Mahalik et al., 2003; Mahalik, Morray, et al., 2005; Parent & Moradi, 2009, 2010). When subscale marker items were selected to form abbreviated 11- and 22-item versions of the CMNI in past research, investigators reported problematically low reliability (e.g., Cronbach's alphas of 0.65 and 0.70 for the CMNI-22 [Burns & Mahalik, 2008; Rochlen, McKelley, Suizzo, & Scaringi, 2008] and theta reliability coefficient of .64 for the CMNI-11 [Mahalik et al., 2007]), challenging the unidimensionality of these selected items. In light of such data, and because the theoretical basis and aim of the CMNI and CFNI were to reflect multiple dimensions of gender norms, it seems more appropriate to examine relations with specific dimensions of conformity to masculine and feminine norms using subscale scores rather than overall scale scores. Use of the CMNI-46 and CFNI-45 can facilitate efficiency in such use. Finally, conformity to masculine norms is typically studied with men, and conformity to feminine norms is typically studied with women. However, conformity or nonconformity to masculine and feminine norms may be relevant to all gender groups, including transgender individuals.

Thus, research is needed to examine the applicability and psychometric properties of these measures across gender groups.

Identity Status Attitudes

Models of gender-related identity status attitudes generally aim to capture connections between personal and collective aspects of identity for women and men. Such models have received greater attention in research with women than with men. Measures for women and men also emerged from different theoretical foundations, including models of feminist and womanist identity for women and male reference group identity dependence for men. Some measures are also emerging to capture intersecting identity consciousness and transgeder identity variables.

Gender-related identity. Feminist and womanist identity development models drew from literature on women's experiences, gender identity, and Black racial identity (Cross, 1971; Helms, 1984) to capture movement from low to high levels of personal and collective consciousness and empowerment. Developmental assumptions of both models have been softened and the models are currently operationalized as a set of nonlinear attitudes that reflect profiles rather than stages of feminist or womanist identity attitudes (for reviews, see Moradi, 2005; Moradi, Subich, & Phillips, 2002a; Moradi, Yoder, & Berendsen, 2004).

Feminist identity development comprises five identities: (a) passive acceptance, or unexamined acceptance of traditional gender roles and denial of sexism; (b) revelation, or increased realization about sexism, usually accompanied by anger about societal sexism and guilt about one's own participation in sexism; (c) embeddedness and emanation, idealizing women and women's culture; (d) synthesis of feminist consciousness with other aspects of a positive self-concept and an individual differences approach (rather than dichotomous thinking) toward women and men; and (e) active commitment to societal change and eliminating oppression (Downing & Roush, 1985). This model was initially operationalized with the 37-item Feminist Identity Scale (FIS; Rickard, 1989) and the 39-item Feminist Identity Development Scale (FIDS; Bargad & Hyde, 1991). Fischer and colleagues (2000) developed the 33-item Feminist Identity Composite (FIC) by integrating items from the aforementioned FIS and FIDS to capitalize on their psychometric strengths.

Psychometric strengths and limitations have been noted for these three instruments (see Moradi & Subich, 2002a; Moradi et al., 2002a). Specifically, Cronbach's alphas have been acceptable for some but below 0.70 for other subscale items (e.g., FIS Passive Acceptance and Embeddedness Emanation and FIDS Revelation and Synthesis), confirmatory factor analyses raised questions about the structural properties of all three measures and some scholars have questioned the conceptual and empirical value of Synthesis (Liss & Erchull, 2010). Nevertheless, use of these measures in a large body of research has produced validity-consistent data, linking feminist identity attitudes with variables such as activism in women's organizations, egalitarian attitudes, perceptions of sexism, empowerment, psychological well-being, and psychological distress (e.g., Moradi & Subich, 2002a, 2002b; Peterson, Grippo, & Tantleff-Dunn, 2008; Saunders & Kashubeck-West, 2006; White, Strube, & Fisher, 1998; Yoder, Perry, & Saal, 2007).

The womanist identity development model describes moving from an externally based sociocultural or sociopolitical definition of oneself as a woman to an internally based self-definition, without a focus on feminist identification or activism (Carter & Parks, 1996; Ossana, Helms, & Leonard, 1992). This process is posited to be similar across women of diverse ethnic, racial, social class, and other backgrounds. Womanist identity development consists of four statuses: (a) Preencounter, reflecting a denial of sexism and rigid conformity to societal values that privilege men relative to women; (b) Encounter with new experiences that challenge Preencounter values, promote awareness of sexism, increase identification with womanhood, and foster exploration of alternative roles for women and men; (c) Immersion–Emersion, involving the idealization of women, rejection of patriarchal definitions, and search for positive definitions of womanhood; and (d) Internalization of a personally defined positive view of womanhood into one's identity (Carter &

Parks, 1996; Ossana et al., 1992). This model was operationalized with the 43-item Womanist Identity Attitudes Scale (WIAS; Ossana, 1986; Ossana et al., 1992).

Exploratory and confirmatory factor analyses have raised questions about the fit of WIAS data with the hypothesized model and Cronbach's alpha for WIAS subscale items have been variable, ranging from the 0.30s to 0.50s for Preencounter and Encounter and from the 0.30s to 0.80s for Immersion–Emersion and Internalization (see Moradi et al., 2004). Nevertheless, as with feminist identity development measures, the WIAS has yielded some theoretically consistent findings including expected subscale associations with self-esteem, self-efficacy, and external locus of control as well as similarities in means and intercorrelations between Black and White women (e.g., Boisnier, 2003; Ossana et al., 1992; for a review, see Moradi, 2005). However, links between WIAS scores and gender-related attitudes have been mixed. As expected, Preencounter scores were correlated negatively with egalitarian attitudes and positively with sexist attitudes, and Internalization scores were correlated positively with egalitarian attitudes; but, contrary to their conceptual definitions, Encounter and Immersion–Emersion scores were generally uncorrelated with egalitarian and sexist attitudes (Moradi et al. 2004).

Focusing on men's identity, the male reference group identity dependence model reflects the centrality of men as a reference group to individual men's self-concept (Wade, 1998). Wade (1998) originally proposed three statuses: (a) No Reference Group reflects an undefined gender role self-concept characterized by lack of connection or perceived similarity to other men; (b) Reference Group Dependent reflects rigid conformity to externally defined gender role self-concept, identifying with some but not other groups of men; and (c) Reference Group Nondependent reflects a flexible and internally defined gender role self-concept and sense of commonality with various groups of men. On the basis of factor analyses, Wade and Gelso (1998) divided Reference Group Nondependent into a Similarity dimension (feeling connected with all men) and a Diversity dimension (appreciation of differences among men). This model was operationalized with the 30-item Reference Group Identity Dependence Scale (RGIDS; Wade & Gelso, 1998).

Findings of exploratory factor analyses have been consistent with the posited four-factor structure with predominantly White men (Wade & Gelso, 1998), but suggested an alternative three-factor structure with African American men, with the factors reflecting (a) understanding of and connection with diversity among men, (b) disconnection from other men, and (c) connection with some men but not others (Wade, 2008). Cronbach's alphas for subscale items (according to the four or three factor solutions) have generally ranged in the .70 and .80s with the exception of values mostly in the .60s for Reference Group Dependence (Wade, 2008; Wade & Brittan-Powell, 2001; Wade & Donis, 2007; Wade & Gelso, 1998). Research using the RGIDS has yielded some validity-consistent links with psychological distress, well-being, ego identity statuses, instrumentality and expressiveness, perceived romantic relationship quality, and masculine gender role conflict (Wade, 2008; Wade & Brittan-Powell, 2001; Wade & Donis, 2007; Wade & Gelso, 1998). However, associations with traditional masculinity ideology have varied in direction, magnitude, and significance for No Reference Group and Reference Group Non-Dependent Diversity attitudes (Wade, 2008; Wade & Brittan-Powell, 2001; Wade & Donis, 2007), raising questions about the posited role of masculinity ideology in these identity statuses.

Identity intersections. Two measures are selected to exemplify different foci in operationalizing gender-related identity intersections. One approach is to directly assess attitudes reflecting an intersectionality consciousness or the perspective that gender and other dimensions of identity are connected in one's experiences. Another approach is to assess gender-related identity from the perspective of a social group with intersecting identities (e.g., lesbian women).

An exemplar of the intersectionality consciousness approach is the 14-item Womanist Consciousness Scale (WCS; King, 2003) designed to measure the fusion of race and gender in the identity of women of color. Psychometric evidence for the WCS is limited but promising. Specifically, in a sample of African American college women, Cronbach's

alpha was 0.86, and WCS scores were related positively with ethnic and feminist consciousness and with attributing a negative experience to fused ethnic and gender discrimination (King, 2003).

Another approach to operationalizing intersecting identities is reflected in McCarn and Fassinger's (1996) model of identity formation processes for lesbian women. This model teases apart individual sexual identity formation from group and sociopolitical identity formation along four, not necessarily linear or parallel, phases: (a) Awareness of feeling different from the heterosexual norm (individual) and recognizing the nonuniversality of heterosexuality (group); (b) Exploration of erotic feelings toward women (individual) and one's position among and attitudes toward sexual minority people and communities (group); (c) Deepening/Commitment to one's sexuality and sexual identity (individual) and shared experiences with sexual minority communities (group); (d) and Internalization/Synthesis of one's sexual identity (individual) and group membership identity (group). This model was operationalized with the 40-item Lesbian Identity Questionnaire (LIQ; see Swann & Spivey, 2004; Tomlinson & Fassinger, 2003) and served as a basis for similar measures for use with other sexual identity groups (e.g., Fassinger & Miller 1996; Worthington, Navarro, Savoy, & Hampton, 2008).

Psychometric data about the LIQ are limited, but some evidence of the validity of LIQ scores is available. For example, consistent with the posited distinctiveness of individual and group levels of lesbian identity formation, self-esteem was correlated negatively with Awareness, Exploration, and Deepening/Commitment phases of group identity, self-esteem was uncorrelated with these phases of individual identity, and self-esteem and indicators of vocational maturity were correlated positively with the Synthesis/Internalization phase of both levels of identity (Swann & Spivey, 2004; Tomlinson & Fassinger, 2003). These findings suggest that high awareness, exploration, and commitment to a stigmatized sociopolitical identity may be associated with self-esteem costs, whereas individual and group level identity synthesis and internalization may be associated with self-esteem and vocational development benefits.

Transgender identity. As with research on identity intersections, limited research has been devoted to assessing transgender identity; two approaches are described here as potentially promising. First, the Transgender Adaptation and Integration Measure (TG AIM; Sjoberg, Walch, & Stanny, 2006) is a 15-item measure developed using rational instrument construction procedures (e.g., literature review, expert review) followed by empirical evaluation of factor structure, reliability, and validity. Sjoberg et al.'s (2006) exploratory factor analyses suggested four factors reflecting (a) fears about social rejection and discrimination (Fears), (b) intrapersonal distress (Distress), (c) coping and gender-reorientation behaviors (Coping), and (d) internality and importance of transgender identity (Internality); the fourth factor was dropped because of reliability and validity concerns. The first three factors' items yielded Cronbach's alphas over 0.70. Correlations among the three retained factors were consistent with validity, that is, Fears was correlated positively with Distress and with Coping, and Distress and Coping were correlated negatively with one another. Furthermore, greater Fears and Distress were correlated with reports of lower quality of life and self-esteem, and with greater psychological symptomatology (Sánchez & Vilain, 2009; Sjoberg et al., 2006).

The 16-item Collective Self-Esteem Scale (CSES; Luhtanen & Crocker, 1992) also has been used to operationalize aspects of collective identity among male-to-female transsexual individuals (Sánchez & Vilain, 2009). The CSES assesses membership esteem (i.e., individuals' judgments of how good or worthy they are as members of their social group), private esteem (i.e., individuals' judgments of how good their social groups are), public esteem (i.e., individuals' judgments of how other people view their groups), and identity esteem (i.e., the importance of group membership to individuals' self-concepts) and has been used with many different groups including women and men (e.g., Burn, Aboud, & Moyles, 2000). In Sánchez and Vilain's (2009) sample of transsexual individuals, Cronbach's alphas for Membership, Private, Public, and Identity esteem subscale items ranged from 0.70 to 0.83. These dimensions of collective self-esteem were correlated positively with one another and

correlated negatively with TG AIM Fear and Distress scores and with psychological symptomatology. These findings provide preliminary support for use of the CSES with transgender populations.

Areas for advancement. A strength of the literature on some of the identity development models reviewed here is that ongoing research on the psychometric properties and conceptual underpinnings of the measures is informing model and measurement refinements; such a recursive process is particularly evident with feminist identity development measures. However, psychometric advances continue to be needed with regard to the reliability and structural validity of all of these measures. Areas of overlap and distinctiveness among models are important to clarify. For example, the womanist and feminist identity development models both involve moving from denial to acknowledgment of women's oppression and exploring, synthesizing, and internalizing alternatives to oppressive roles. Indeed, among African American and White women, moderate to high positive correlations were found between parallel womanist and feminist identity development attitudes (Boisnier, 2003; Hoffman, 2006). Thus, efforts are needed to consolidate overlapping aspects of these models and measures and to draw out their unique contributions. Relatedly, in a sample of women of various racial/ethnic backgrounds, feminist and womanist identity attitudes that reflect heightened awareness of gender oppression (i.e., FIDS Revelation, Embeddeness–Emanation, Active Commitment; WIAS Immersion–Emersion) were associated with greater exploration of and commitment to ethnic identity (Hoffman, 2006) suggesting the importance of attending to intersections of women's gender and ethnic identities (e.g., Moradi, Subich, & Phillips, 2002b; Vandiver, 2002). It is important to note that the developmental roots of feminist and womanist identity models have been circumvented in operationalizing them; thus, research is needed to directly examine posited developmental processes to inform theory and measurement refinement. The potential for models and measures that capture statuses of feminist consciousness for men also remains underdeveloped.

Similarly, the male reference group dependence model and scale raise the possibility of applying reference group models to understanding other gender identities as well (e.g., women, transgender). One potential for reference group models is the integration of multiple possible reference groups. For example, assessing directly what reference groups are most salient for respondents could provide an avenue for capturing identity intersections (e.g., gender, ethnicity, race, sexual orientation). However, the central tenet regarding the importance of men as reference group to self-concept across the identity statuses remains to be examined and would need clarification in extensions of the model to other gender groups as well. As an example, investigating the associations of reference group statuses with dimensions of collective self-esteem would be a fruitful approach to evaluating this aspect of model and measurement validity.

Perceived Experiences of Sexism and Gender-Based Differential Treatment

Landrine and Klonoff (1997) defined *sexist events* as "discriminatory acts or events that happen to women because they are women" (p. 22). Sexist events differ from gender-based differential treatment directed at men in that the former reflect and reinforce societal power hierarchies that accord women subordinate status compared with men (Krieger, 1999; Major & O'Brien, 2005). This view is consistent with findings that negative treatment based on one's group status has different consequences for high- versus low-status people (e.g., Foster, Arnt, & Honkola, 2004; Major & O'Brien, 2005; Schmitt & Branscombe, 2002) and that gender-based discrimination has different correlates and outcomes for women and men (e.g., Kobrynowicz & Branscombe, 1997; Schmitt, Branscombe, Kobrynowicz, & Owen, 2002).

The Schedule of Sexist Events (SSE; Klonoff & Landrine, 1995) is among the most frequently used measures of women's perceived experiences of sexism. Informed by the daily hassles and stressful life events literatures, the 20-item SSE was designed to assess self-reported frequency of lifetime and recent (i.e., past year) sexist events and the perceived stressfulness of those events (Klonoff & Landrine, 1995; Landrine & Klonoff, 1997). Across diverse samples of women, Recent, Lifetime, and Appraisal

items have generally yielded Cronbach's alphas in the .90s (e.g., DeBlaere & Moradi, 2008; Fischer et al., 2000; Klonoff & Landrine, 1995; Landrine & Klonoff, 1997; Matteson & Moradi, 2005; Moradi & Subich, 2003, 2004; Szymanski, 2005).

With regard to validity, SSE scores have had nonsignificant or small correlations with social desirability and positive correlations with such variables as psychological distress, involvement in women's organizations, and awareness of sexism (e.g., Fischer et al., 2000; Moradi & Funderburk, 2006; Moradi & Subich, 2002b, 2003; Szymanski, 2005). Exploratory and confirmatory factor analyses of data from White and African American/Black women suggested that, in addition to overall lifetime and recent frequency and appraisal scores, the SSE can be scored to assess lifetime experiences and appraisal of (a) intimate/personal sexist events and (b) unfair treatment in public contexts; and recent experiences of (a) sexist degradation and its consequences, (b) unfair or sexist events at work/school, and (c) unfair treatment in distant and close relationship contexts (DeBlaere & Moradi, 2008; Klonoff & Landrine, 1995; Matteson & Moradi, 2005). However, substantial overlap among Recent, Lifetime, and Appraisal scores has led some researchers to use only SSE-Recent (e.g., Moradi & Subich, 2003; Szymanski, 2005).

Whereas the SSE is designed specifically for use with women, Swim, Cohen, and Hyers (1998) developed a measure for use with women and men. Swim et al.'s (1998) measure was based on diary data from college women and men who recorded their observations of gender-related incidents over a 2-week period. On the basis of these data, a 25-item self-report measure was developed to assess (a) sexual objectification directed at the respondent, (b) gender role stereotyping directed at the respondent, (c) gender role stereotyping of women in general, and (d) gender role stereotyping of men in general (Swim et al., 1998; Swim, Hyers, Cohen, & Ferguson, 2001). College women reported more gender-related incidents directed at themselves and their own gender group than did men. Also, both women and men reported more gender-related incidents directed at women in general than against men in general. Beyond these initial data, psychometric evidence regarding Swim et al.'s (1998) measure is limited, although the subscale measuring sexual objectification has been used with college women (Mitchell & Mazzeo, 2009; Moradi, Dirks, & Matteson, 2005) and sexual minority men (Wiseman & Moradi, 2010).

Areas for advancement. Although ample psychometric evidence exists for the SSE across diverse samples of women, further psychometric investigation of Swim et al.'s (1998) measure is needed (for a review, see Moradi & DeBlaere, 2010). Such research may point to areas of overlap and distinction between the two measures. An important consideration in investigating the reliability and factor structure of these measures is the extent to which the events assessed can be thought of as causal indicators that do not necessarily covary but cumulatively shape the construct of interest or as effect indicators that covary and reflect a unidimensional underlying construct (Moradi & DeBlaere, 2010; Streiner, 2003). The daily diary root of Swim et al.'s measure also is a reminder of the utility of experience sampling in assessing gender-related experiences. Experience sampling can be used to minimize recall effects and to examine how individual differences and contexts shape intraindividual variability in experiences of sexist events (Scollon, Kim-Prieto, & Diener, 2003). Swim et al.'s (1998) measure also can be useful in elucidating the correlates of men's perceived experiences of gender-based differential treatment.

CONCLUSION

Three decades ago, Beere (1979) noted that construct ambiguity, proliferation of measures, and limited psychometric support plagued assessment of gender-related constructs. As evident in this review, these concerns persist for some gender-related constructs, but research also is evolving toward greater construct specificity and psychometric sophistication. As an echo of Beere's appraisal and a thread across the areas for advancement noted throughout this review, it seems fruitful to redirect efforts from measurement proliferation to measurement consolidation and refinement. There is useful groundwork

for such efforts in many classes of measures. For instance, investigating redundancy and uniqueness among masculine ideology measures, feminine ideology measures, masculine stress/conflict and norm conformity measures, and feminist and womanist identity measures can guide the clarification, consolidation, and improvement of measurement within and across classes of gender-related constructs.

There are, however, areas where construct definition and measurement development are in more nascent stages and warrant greater attention. One such area is integration of transgender issues into assessment of gender-related constructs; the measures included in this review suggest promising starts to this area of investigation. Operationalizing the complexities of intersecting identities is also an important direction for further investigation. In our view, transgender issues and intersecting identities represent two important leading edges in the assessment of gender-related constructs. The conceptual and psychometric advances reflected in the rich research on gender-related constructs can inform these next important steps in the evolution of this literature.

References

American Psychological Association Division 44/Committee on Lesbian, Gay, and Bisexual Concerns Joint Task Force. (2000). Guidelines for psychotherapy with lesbian, gay, and bisexual clients. *American Psychologist, 55*, 1440–1451.

American Psychological Association Task Force on Appropriate Therapeutic Responses to Sexual Orientation. (2009). *Report of the Task Force on Appropriate Therapeutic Responses to Sexual Orientation*. Washington, DC: American Psychological Association.

American Psychological Association Task Force on Gender Identity and Gender Variance. (2008). *Report of the Task Force on Gender Identity and Gender Variance*. Washington, DC: American Psychological Association.

Archer, J. (1989). The relationship between gender-role measures: A review. *British Journal of Social Psychology, 28*, 173–184.

Bargad, A., & Hyde, J. S. (1991). Women's studies: A study of feminist identity development in women. *Psychology of Women Quarterly, 15*, 181–201. doi:10.1111/j.1471-6402.1991.tb00791.x

Beere, C. A. (1990). *Gender roles: A handbook of tests and measurements*. New York, NY: Greenwood Press.

Beere, C. A., King, D. W., Beere, D. B., & King, L. A. (1984). The Sex-Role Egalitarianism Scale: A measure of attitudes toward equality between the sexes. *Sex Roles, 10*, 563–576. doi:10.1007/BF00287265

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*, 155–162. doi:10.1037/h0036215

Bem, S. L. (1977). On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology, 45*, 196–205. doi:10.1037/0022-006X.45.2.196

Bem, S. L. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. New Haven, CT: Yale.

Berger, J. M., Levant, R. F., McMillan, K. K., Kelleher, W., & Sellers, A. (2005). Impact of gender role conflict, traditional masculinity ideology, alexithymia, and age on men's attitudes toward psychological help seeking. *Psychology of Men and Masculinity, 6*, 73–78. doi:10.1037/1524-9220.6.1.73

Betz, N. E., & Fitzgerald, L. F. (1993). Individuality and diversity: Theory and research in counseling psychology. *Annual Review of Psychology, 44*, 343–381. doi:10.1146/annurev.ps.44.020193.002015

Boisnier, A. D. (2003). Race and women's identity development: Distinguishing between feminism and womanism among Black and White women. *Sex Roles, 49*, 211–218. doi:10.1023/A:1024696022407

Bornstein, K. (1998). *My gender workbook*. New York, NY: Routledge.

Boyraz, G., & Sayger, T. V. (2009). An exploratory path analysis of the factors contributing to life satisfaction in fathers. *Journal of Positive Psychology, 4*, 145–154. doi:10.1080/17439760802650592

Brannon, R., & Juni, S. (1984). A scale for measuring attitudes about masculinity. *Psychological Documents, 14,* 2612.

Burn, S. M., Aboud, R., & Moyles, C. (2000). The relationship between gender social identity and support for feminism. *Sex Roles, 42*, 1081–1089. doi:10.1023/A:1007044802798

Burn, S. M., & Ward, A. Z. (2005). Men's conformity to traditional masculinity and relationship satisfaction. *Psychology of Men and Masculinity, 6*, 254–263. doi:10.1037/1524-9220.6.4.254

Burns, S. M., & Mahalik, J. R. (2008). Sexual functioning as a moderator of the relationship between masculinity and men's adjustment following treatment for prostate cancer. *American Journal of Men's Health, 2*, 6–16. doi:10.1177/1557988307304325

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized*

Minnesota Multiphasic Personality Inventory: MMPI–2. Minneapolis: University of Minnesota Press.

Carter, R. T., & Parks, E. E. (1996). Womanist identity and mental health. *Journal of Counseling and Development, 74*, 484–489. doi:10.1002/j.1556-6676.1996.tb01897.x

Choi, N., & Fuqua, D. R. (2003). The structure of the Bem Sex Role Inventory: A summary report of 23 validation studies. *Educational and Psychological Measurement, 63*, 872–887. doi:10.1177/0013164403258235

Collins, P. H. (1990). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Boston, MA: Unwin Hyman.

Cross, W. E. (1971). The Negro-to-Black conversion experience: Toward a psychology of Black liberation. *Black World, 20*, 13–27.

DeBlaere, C., & Moradi, B. (2008). Structures of the Schedules of Racist and Sexist Events: Confirmatory factor analysis of African American women's responses. *Psychology of Women Quarterly, 32*, 83–94. doi:10.1111/j.1471-6402.2007.00409.x

Downing, N. E., & Roush, K. L. (1985). From passive acceptance to active commitment: A model of feminist identity development for women. *The Counseling Psychologist, 13*, 695–709. doi:10.1177/0011000085134013

Eisler, R. M., & Skidmore, J. R. (1987). Masculine gender role stress: Scale development and component factors in the appraisal of stressful situations. *Behavior Modification, 11*, 123–136. doi:10.1177/01454455870112001

Fassinger, R. E. (1994). Development and testing of the Attitudes Toward Feminism and the Women's Movement (FWM) Scale. *Psychology of Women Quarterly, 18*, 389–402. doi:10.1111/j.1471-6402.1994.tb00462.x

Fassinger, R. E., & Miller, B. A. (1997). Validation of an inclusive model of sexual minority identity formation on a sample of gay men. *Journal of Homosexuality, 32*, 53–78. doi:10.1300/J082v32n02_04

Fausto-Sterling, A. (1993). The five sexes: Why male and female are not enough. *The Sciences, 33*, 20–25.

Fischer, A. R., Tokar, D. M., Good, G. E., & Snell, A. F. (1998). More on the structure of male role norms: Exploratory and multiple sample confirmatory analyses. *Psychology of Women Quarterly, 22*, 135–155. doi:10.1111/j.1471-6402.1998.tb00147.x

Fischer, A. R., Tokar, D. M., Mergl, M. M., Good, G. E., Hill, M. S., & Blum, S. A. (2000). Assessing women's feminist identity development: Studies of convergent, discriminant, and structural validity. *Psychology of Women Quarterly, 24*, 15–29. doi:10.1111/j.1471-6402.2000.tb01018.x

Foster, M. D., Arnt, S., & Honkola, J. (2004). When the advantaged become disadvantaged: Men's and women's actions against gender discrimination. *Sex Roles, 50*, 27–36. doi:10.1023/B:SERS.0000011070.24600.92

Fox, R. E. (1988). Proceedings of the American Psychological Association, Incorporated, for the year 1987: Minutes of the annual meeting of the Council of Representatives August 27 and 30, 1987, New York, and February 5–7, 1988, Washington, DC. *American Psychologist, 43*, 508–531. doi:10.1037/h0091999

Gillen, M. M., & Lefkowitz, E. S. (2006). Gender role development and body image among male and female first year college students. *Sex Roles, 55*, 25–37. doi:10.1007/s11199-006-9057-4

Gillespie, B. L., & Eisler, R. M. (1992). Development of the feminine gender role stress scale: A cognitive–behavioral measure of stress, appraisal, and coping for women. *Behavior Modification, 16*, 426–438. doi:10.1177/01454455920163008

Glick, P., Fiske, S., Mladinic, A., Saiz, J., Abrams, D., Masser, B., . . . Lopez, W. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*, 763–775. doi:10.1037/0022-3514.79.5.763

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*, 491–512. doi:10.1037/0022-3514.70.3.491

Glick, P., & Fiske, S. T. (1999). The Ambivalence Toward Men Inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of Women Quarterly, 23*, 519–536. doi:10.1111/j.1471-6402.1999.tb00379.x

Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., . . . Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, *86*, 713–728. doi:10.1037/0022-3514.86.5.713

Goodman, M. B., & Moradi, B. (2008). Attitudes and behaviors toward lesbian and gay persons: Critical correlates and mediated relations. *Journal of Counseling Psychology, 55*, 371–384. doi:10.1037/0022-0167.55.3.371

Hathaway, S. R., & McKinley, J. C. (1940). *The MMPI manual*. New York, NY: Psychological Corporation.

Helms, J. E. (1984). Toward a theoretical explanation of the effects of race on counseling: A Black and White model. *The Counseling Psychologist, 12*, 153–165. doi:10.1177/0011000084124013

Hill, D. B., & Willoughby, B. L. B. (2005). The development and validation of the Genderism and Transphobia Scale. *Sex Roles, 53*, 531–544. doi:10.1007/s11199-005-7140-x

Hoffman, R. M. (2006). Gender self-definition and gender self-acceptance in women: Intersections

with feminist, womanist, and ethnic identities. *Journal of Counseling and Development, 84*, 358–372. doi:10.1002/j.1556-6678.2006.tb00415.x

Hoffman, R. M., & Borders, D. (2001). Twenty-five years after the Bem Sex-Role Inventory: A reassessment and new issues regarding classification variability. *Measurement and Evaluation in Counseling and Development, 34*, 39–55.

Holz, K., & DiLalla, D. L. (2007). Men's fear of unintentional rape: Measure development and psychometric evaluation. *Psychology of Men and Masculinity, 8*, 201–214. doi:10.1037/1524-9220.8.4.201

Hurt, M. M., Nelson, J. A., Turner, D. L., Haines, M. E., Ramsey, L. R., Erchull, M. J., . . . Liss, M. (2007). Feminism: What is it good for? Feminine norms and objectification as the link between feminist identity and clinically relevant outcomes. *Sex Roles*, *57*, 355–363. doi:10.1007/s11199-007-9272-7

Jakupcak, M., Tull, M. T., & Roemer, L. (2005). Masculinity, shame, and fear of emotions as predictors of men's expressions of anger and hostility. *Psychology of Men and Masculinity, 6*, 275–284. doi:10.1037/1524-9220.6.4.275

Johnson, A. G. (2006). *Privilege, power, and difference* (2nd ed.). New York, NY: McGraw Hill.

King, K. R. (2003). Do you see what I see? Effects of group consciousness on African American women's attributions to prejudice. *Psychology of Women Quarterly, 27*, 17–30. doi:10.1111/1471-6402.t01-2-00003

King, L. A., & King, D. W. (1997). Sex-role egalitarianism: Development, psychometric properties, and recommendations for future research. *Psychology of Women Quarterly, 21*, 71–87. doi:10.1111/j.1471-6402.1997.tb00101.x

King, L. A., King, D. W., Gudanowski, D. M., & Taft, C. T. (1997). Latent structure of the sex-role egalitarianism scale: Confirmatory factor analyses. *Sex Roles, 36*, 221–234. doi:10.1007/BF02766269

Klonoff, E. A., & Landrine, H. (1995). The Schedule of Sexist Events: A measure of lifetime and recent sexist discrimination in women's lives. *Psychology of Women Quarterly, 19*, 439–470. doi:10.1111/j.1471-6402.1995.tb00086.x

Kobrynowicz, D., & Branscombe, N. R. (1997). Who considers themselves victims of discrimination? Individual difference predictors of perceived gender discrimination in women and men. *Psychology of Women Quarterly, 21*, 347–363. doi:10.1111/j.1471-6402.1997.tb00118.x

Krieger, N. (1999). Embodying inequality: A review of concepts, measures, and methods for studying health consequences of discrimination. *International Journal of Health Services, 29*, 295–352. doi:10.2190/M11W-VWXE-KQM9-G97Q

Landrine, H., & Klonoff, E. A. (1997). *Discrimination against women: Prevalence, consequences, remedies.* Thousand Oaks, CA: Sage.

Lease, S., Çiftçi, A., Demir, A., & Boyraz, G. (2009). Structural Validity of Turkish Versions of the Gender Role Conflict Scale and Male Role Norms Scale. *Psychology of Men and Masculinity, 10*, 273–287. doi:10.1037/a0017044

Lefkowitz, E. S., Shearer, C. L., Gillen, M. M., & Espinosa-Hernandez, G. (2006). *Measuring female role norms in male and female emerging adults.* Unpublished manuscript.

Levant, R. F., & Fischer, J. (1998). The Male Role Norms Inventory. In C. M. Davis, W. H. Yarber, R. Bauserman, G. Schreer, & S. L. Davis (Eds.), *Handbook of sexuality-related measures* (pp. 469–472). Thousand Oaks, CA: Sage.

Levant, R. F., Hirsch, L., Celentano, E., Cozza, T., Hill, S., & MacEachern, M. (1992). The male role: An investigation of norms and stereotypes. *Journal of Mental Health Counseling, 14*, 325–337.

Levant, R. F., Rankin, T. J., Williams, C. M., Hasan, N. T., & Smalley, K. B. (2010). Evaluation of the Factor Structure and Construct Validity of Scores on the Male Role Norms Inventory—Revised (MRNI–R). *Psychology of Men and Masculinity, 11*, 25–37. doi:10.1037/a0017637

Levant, R. F., & Richmond, K. (2007). A review of research in masculinity ideologies using the Male Role Norms Inventory. *Journal of Men's Studies, 15*, 130–146. doi:10.3149/jms.1502.130

Levant, R. F., Richmond, K., Cook, S., House, A. T., & Aupont, M. (2007). The Femininity Ideology Scale: Factor structure, reliability, convergent and discriminant validity, and social contextual variation. *Sex Roles, 57*, 373–383. doi:10.1007/s11199-007-9258-5

Levant, R. F., Smalley, K. B., Aupont, M., House, A., Richmond, K., &. Noronha, D. (2007). Validation of the Male Role Norms Inventory—Revised. *Journal of Men's Studies*, *15*, 83–100. doi:10.3149/jms.1501.83

Lewin, M., & Wild, C. L. (1991). The impact of the feminist critique on tests, assessment, and methodology. *Psychology of Women Quarterly, 15*, 581–596. doi:10.1111/j.1471-6402.1991.tb00432.x

Liss, M., & Erchull, M. J. (2010). Everyone feels empowered: Understanding feminist self-labeling. *Psychology of Women Quarterly, 34*, 85–96. doi:10.1111/j.1471-6402.2009.01544.x

Locke, B. D., & Mahalik, J. R. (2005). Examining masculine norms, problem drinking, and athletic involvement as predictors of sexual aggression in college men. *Journal of Counseling Psychology, 52*, 279–283. doi:10.1037/0022-0167.52.3.279

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity.

Personality and Social Psychology Bulletin, *18*, 302–318. doi:10.1177/0146167292183006

Mahalik, J., Aldarondo, E., Gilbert-Gokhale, S., & Shore, E. (2005). The role of insecure attachment and gender role stress in predicting controlling behavior in men who batter. *Journal of Interpersonal Violence, 20*, 617–631. doi:10.1177/0886260504269688

Mahalik, J. R., Burns, S. M., & Syzdek, M. (2007). Masculinity and perceived normative health behaviors as predictors of men's health behaviors. *Social Science and Medicine, 64*, 2201–2209. doi:10.1016/j.socscimed.2007.02.035

Mahalik, J. R., Locke, B., Ludlow, L., Diemer, M., Scott, R. P. J., Gottfried, M., . . . Freitas, G. (2003). Development of the Conformity to Masculine Norms Inventory. *Psychology of Men and Masculinity*, *4*, 3–25. doi:10.1037/1524-9220.4.1.3

Mahalik, J. R., Morray, E. B., Coonerty-Femaino, A., Ludlow, L. H., Slattery, S. M., & Smiler, A. (2005). Development of the Conformity to Feminine Norms Inventory. *Sex Roles, 52*, 417–435. doi:10.1007/s11199-005-3709-7

Mahalik, J. R., & Rochlen, A. B. (2006). Men's likely responses to clinical depression: What are they and do masculinity norms predict them? *Sex Roles, 55*, 659–667. doi:10.1007/s11199-006-9121-0

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology, 56*, 393–421. doi:10.1146/annurev.psych.56.091103.070137

Matteson, A. V., & Moradi, B. (2005). Examining the structure of the schedule of sexist events: A replication and extension. *Psychology of Women Quarterly, 29*, 47–57. doi:10.1111/j.0361-6843.2005.00167.x

McCarn, S. R., & Fassinger, R. E. (1996). Revisioning sexual minority identity formation: A new model of lesbian identity and its implications for counseling and research. *The Counseling Psychologist, 24*, 508–534. doi:10.1177/0011000096243011

McCreary, D. R., Newcomb, M. D., & Sadava, S. W. (1999). The male role, alcohol use, and alcohol problems: A structural modeling examination in adult women and men. *Journal of Counseling Psychology, 46*, 109–124. doi:10.1037/0022-0167.46.1.109

Mitchell, K. S., & Mazzeo, S. E. (2009). Evaluation of a structural model of objectification theory and eating disorder symptomatology among European American and African American undergraduate women. *Psychology of Women Quarterly, 33*, 384–395. doi:10.1111/j.1471-6402.2009.01516.x

Moore, T. M., & Stuart, G. L. (2005). A review of the literature on masculinity and partner violence. *Psychology of Men and Masculinity, 6*, 46–61. doi:10.1037/1524-9220.6.1.46

Moradi, B. (2005). Advancing womanist identity development: Where we are and where we need to go. *The Counseling Psychologist, 33*, 225–253. doi:10.1177/0011000004265676

Moradi, B., & DeBlaere, C. (2010). Women's experiences of sexist discrimination: Review of research and directions for centralizing race, ethnicity, and culture. In H. Landrine & N. F. Russo (Eds.), *Handbook of Diversity in Feminist Psychology* (pp. 173–210). New York, NY: Springer.

Moradi, B., Dirks, D., & Matteson, A. V. (2005). Roles of sexual objectification experiences and internalization of standards of beauty in eating disorder symptomatology: A test and extension of objectification theory. *Journal of Counseling Psychology, 52*, 420–428. doi:10.1037/0022-0167.52.3.420

Moradi, B., & Funderburk, J. R. (2006). Roles of perceived sexist events and perceived social support in the mental health of women seeking counseling. *Journal of Counseling Psychology, 53*, 464–473. doi:10.1037/0022-0167.53.4.464

Moradi, B., & Subich, L. M. (2002a). Feminist identity development measures: Comparing the psychometrics of three instruments. *The Counseling Psychologist, 30*, 66–86. doi:10.1177/0011000002301004

Moradi, B., & Subich, L. M. (2002b). Perceived sexist events and feminist identity development attitudes: Links to women's psychological distress. *The Counseling Psychologist, 30*, 44–65. doi:10.1177/0011000002301003

Moradi, B., & Subich, L. M. (2003). A concomitant examination of the relations of perceived racist and sexist events to psychological distress for African American women. *The Counseling Psychologist, 31*, 451–469. doi:10.1177/0011000003031004007

Moradi, B., & Subich, L. M. (2004). Examining the moderating role of self-esteem in the link between experiences of perceived sexist events in psychological distress. *Journal of Counseling Psychology, 51*, 50–56. doi:10.1037/0022-0167.51.1.50

Moradi, B., Subich, L. M., & Phillips, J. (2002a). Beyond revisiting: Moving feminist identity development ahead. *The Counseling Psychologist, 30*, 103–109.

Moradi, B., Subich, L. M., & Phillips, J. (2002b). Revisiting feminist identity development theory, research, and practice. *The Counseling Psychologist, 30*, 6–43. doi:10.1177/0011000002301002

Moradi, B., Yoder, J. D., & Berendsen, L. L. (2004). An evaluation of the psychometric properties of the Womanist Identity Attitudes scale. *Sex Roles, 50*, 253–266. doi:10.1023/B:SERS.0000015556.26966.30

Morawski, J. G. (1985). The measurement of masculinity and femininity: Engendering categorical realities. *Journal of Personality, 53*, 196–223. doi:10.1111/j.1467-6494.1985.tb00364.x

Mussap, A. J. (2008). Masculine gender role stress and the pursuit of muscularity. *International Journal of Men's Health, 7*, 72–89. doi:10.3149/jmh.0701.72

Nagoshi, J. L., Adams, K. A., Terrell, H. K., Hill, E. D., Brzuzy, S., & Nagoshi, C. T. (2008). Gender differences in correlates of homophobia and transphobia. *Sex Roles, 59*, 521–531. doi:10.1007/s11199-008-9458-7

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to measure personality. *Psychological Bulletin*, *92*, 572–580. doi:10.1037/0033-2909.92.3.572

O'Neil, J., Helms, B., Gable, R., David, L., & Wrightsman, L. (1986). Gender Role Conflict Scale: College men's fear of femininity. *Sex Roles, 14*, 335–350. doi:10.1007/BF00287583

O'Neil, J. M. (2008). Men's gender role conflict: 25 year research summary. *The Counseling Psychologist, 36*, 358–445. doi:10.1177/0011000008317057

Ossana, S. M. (1986). *The relationship between women's perceptions of the campus environment and self-esteem as moderated by women's identity attitudes* (Unpublished master's thesis). University of Maryland, College Park.

Ossana, S. M., Helms, J. E., & Leonard, M. M. (1992). Do "womanist" identity attitudes influence college women's self-esteem and perceptions of environmental bias? *Journal of Counseling and Development, 70*, 402–408. doi:10.1002/j.1556-6676.1992.tb01624.x

Parent, M. C., & Moradi, B. (2009). Confirmatory factor analysis of the Conformity to Masculine Norms Inventory and development of the CMNI-46. *Psychology of Men and Masculinity, 10*, 175–189. doi:10.1037/a0015481

Parent, M. C., & Moradi, B. (2010). Confirmatory factor analysis of the Conformity to Feminine Norms Inventory and development of the CFNI-45. *Psychology of Women Quarterly, 34*, 97–109. doi:10.1111/j.1471-6402.2009.01545.x

Parent, M. C., & Moradi, B. (2011a). An abbreviated tool for assessing conformity to masculine norms: Psychometric properties of the Conformity to Masculine Norms Inventory-46. *Psychology of Men & Masculinity, 12,* 339–353.

Parent, M. C., & Moradi, B. (2011b). An abbreviated tool for assessing feminine norm conformity: Psychometric properties of the Conformity to Feminine Norms Inventory-45. *Psychological Assessment, 23,* 958–969.

Parent, M. C., Moradi, B., Rummell, C. M., & Tokar, D. M. (2011). Evidence of construct distinctiveness for conformity to masculine norms. *Psychology of Men & Masculinity, 12,* 354–367.

Parrott, D. J., Peterson, J. L., Vincent, W., & Bakeman, R. (2008). Correlates of anger in response to gay men: Effects of male gender role beliefs, sexual prejudice and masculine gender role stress. *Psychology of Men and Masculinity, 9*, 167–178. doi:10.1037/1524-9220.9.3.167

Pedhazur, E. J., & Tetenbaum, T. J. (1979). Bem Sex Role Inventory: A theoretical and methodological *critique. Journal of Personality and Social Psychology, 37*, 996–1016. doi:10.1037/0022-3514.37.6.996

Peterson, R. D., Grippo, K. P., & Tantleff-Dunn, S. (2008). Empowerment and powerlessness: A closer look at the relationship between feminism, body image and eating disturbance. *Sex Roles, 58*, 639–648. doi:10.1007/s11199-007-9377-z

Petrocelli, J. V. (2002). Ambivalent Sexism Inventory: Where's the ambivalence? *American Psychologist, 57*, 443–444. doi:10.1037/0003-066X.57.6-7.443

Rickard, K. M. (1989). The relationship of self-monitored dating behaviors to level of feminist identity on the feminist identity scale. *Sex Roles, 20*, 213–226. doi:10.1007/BF00287993

Rochlen, A. R., McKelley, R. A., Suizzo, M., & Scaringi, V. (2008). Predictors of relationship satisfaction, psychological well-being, and life satisfaction among an internet sample of stay-at-home fathers. *Psychology of Men and Masculinity, 9*, 17–28. doi:10.1037/1524-9220.9.1.17

Roughgarden, J. (2004). *Evolution's rainbow: Diversity, gender, and sexuality in nature and people.* Berkeley, CA: University of California Press.

Sánchez, F. J., & Vilain, E. (2009). Collective self-esteem as a coping resource for male-to-female transsexuals. *Journal of Counseling Psychology, 56*, 202–209. doi:10.1037/a0014573

Saunders, K. J., & Kashubeck-West, S. (2006). The relations among feminist identity development, gender-role orientation, and psychological well-being in women. *Psychology of Women Quarterly, 30*, 199–211. doi:10.1111/j.1471-6402.2006.00282.x

Schmitt, M. T., & Branscombe, N. R. (2002). The internal and external causal loci of attributions to prejudice. *Personality and Social Psychology Bulletin*, *28*, 620–628. doi:10.1177/0146167202288006

Schmitt, M. T., Branscombe, N. R., Kobrynowicz, D., & Owen, S. (2002). Perceiving discrimination against one's gender group has different implications for well-being in women and men. *Personality and Social Psychology Bulletin*, *28*, 197–210. doi:10.1177/0146167202282006

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies, 4*, 5–34. doi:10.1023/A:1023605205115

Singer, M. I. (1970). Comparison or indicators of homosexuality on the MMPI. *Journal of Consulting*

and Clinical Psychology, 34, 15–18. doi:10.1037/h0028801

Sjoberg, M. D., Walch, S., & Stanny, C. J. (2006). Development and initial psychometric evaluation of the Transgender Adaptation and Integration Measure (TGAIM). *International Journal of Transgenderism, 9*, 35–45. doi:10.1300/J485v09n02_05

Smiler, A. P. (2004). Thirty years after the discovery of gender: Psychological concepts and measures of masculinity. *Sex Roles, 50*, 15–26. doi:10.1023/B:SERS.0000011069.02279.4c

Spence, J. T. (1991). Do the BSRI and PAQ measure the same or different concepts? *Psychology of Women Quarterly, 15*, 141–165. doi:10.1111/j.1471-6402.1991.tb00483.x

Spence, J. T., & Hahn, E. D. (1997). The Attitudes Toward Women Scale and attitude change in college students. *Psychology of Women Quarterly, 21*, 17–34. doi:10.1111/j.1471-6402.1997.tb00098.x

Spence, J. T., & Helmreich, R. L. (1972). The Attitudes Toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology, 2*, 66–67.

Spence, J. T., & Helmreich, R. L. (1978). *Masculinity and femininity: Their psychological dimensions, correlates and antecedents*. Austin: University of Texas Press.

Spence, J. T., Helmreich, R. L., & Stapp, J. (1973). The Personal Attributes Questionnaire: A measure of sex-role stereotypes and masculinity–femininity. *JSAS Catalog of Selected Documents in Psychology, 4*, 43–44.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217–222. doi:10.1207/S15327752JPA8003_01

Swann, S. K., & Spivey, C. A. (2004). The relationship between self-esteem and lesbian identity during adolescence. *Child and Adolescent Social Work Journal, 21*, 629–646. doi:10.1007/s10560-004-6408-2

Swim, J. K., Cohen, L. L., & Hyers, L. L. (1998). Experiencing everyday prejudice and discrimination. In J. K. Swim & C. Stangor (Eds.), *Prejudice: The target's perspective* (pp. 37–60). San Diego, CA: Academic Press.

Swim, J. K., Hyers, L. L., Cohen, L. L., & Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues, 57*, 31–53. doi:10.1111/0022-4537.00200

Szymanski, D. M. (2005). Heterosexism and sexism as correlates of psychological distress in lesbians. *Journal of Counseling and Development, 83*, 355–360. doi:10.1002/j.1556-6678.2005.tb00355.x

Tang, C. S., & Lau, B. H. (1995). The assessment of gender role stress for Chinese. *Sex Roles, 33*, 587–595. doi:10.1007/BF01544682

Tang, C. S., & Lau, B. H. (1996). The Chinese gender role stress scales: Factor structure and predictive validity. *Behavior Modification, 20*, 321–337. doi:10.1177/01454455960203005

Thompson, E., & Pleck, J. H. (1986). The structure of male role norms. *American Behavioral Scientist, 29*, 531–543. doi:10.1177/000276486029005003

Tokar, D. M., Fischer, A. R., Schaub, M., & Moradi, B. (2000). Masculine gender roles and counseling-related variables: Links with and mediation by personality. *Journal of Counseling Psychology, 47*, 380–393. doi:10.1037/0022-0167.47.3.380

Tomlinson, M. J., & Fassinger, R. E. (2003). Career development, lesbian identity development, and campus climate among lesbian college students. *Journal of College Student Development, 44*, 845–860. doi:10.1353/csd.2003.0078

Twenge, G. M. (1997). Attitudes toward women, 1970–1995: A meta-analysis. *Psychology of Women Quarterly, 21*, 35–51. doi:10.1111/j.1471-6402.1997.tb00099.x

Vandiver, B. J. (2002). What do we know and where do we go? *The Counseling Psychologist, 30*, 96–104. doi:10.1177/0011000002301006

van Well, S., Kolk, A. M., & Arrindell, W. A. (2005). Cross-cultural validity of the masculine and feminine gender role stress scales. *Journal of Personality Assessment, 84*, 271–278. doi:10.1207/s15327752jpa8403_06

Vreven, D. L., King, L. A., & King, D. W. (1994). *Item response theory-based information on the Sex-Role Egalitarianism Scale: A technical report to accompany the manual*. Mount Pleasant: Department of Psychology, Central Michigan University.

Wade, J. C. (1998). Male reference group identity dependence: A theory of male identity. *The Counseling Psychologist, 26*, 349–383. doi:10.1177/0011000098263001

Wade, J. C. (2008). Masculinity ideology, male reference group identity dependence, and African American men's health-related attitudes and behaviors. *Psychology of Men and Masculinity*, *9*, 5–16. doi:10.1037/1524-9220.9.1.5

Wade, J. C., & Brittan-Powell, C. S. (2001). Men's attitudes toward race and gender equity: The importance of masculinity ideology, gender-related traits, and reference group identity dependence. *Psychology of Men and Masculinity, 2*, 42–50. doi:10.1037/1524-9220.2.1.42

Wade, J. C., & Donis, E. (2007). Masculinity ideology, male identity, and romantic relationship quality

among heterosexual and gay men. *Sex Roles, 57*, 775–786. doi:10.1007/s11199-007-9303-4

Wade, J. C., & Gelso, C. J. (1998). Reference Group Identity Dependence Scale: A measure of male identity. *The Counseling Psychologist, 26*, 384–412. doi:10.1177/0011000098263002

Ward, L. C., Thorn, B. E., Clements, K. L., Dixon, K. E., & Sanford, S. D. (2006). Measurement of agency, communion, and emotional vulnerability with the Personal Attributes Questionnaire. *Journal of Personality Assessment, 86*, 206–216. doi:10.1207/s15327752jpa8602_10

West, C., & Fenstermaker, S. (1995). Doing difference. *Gender and Society, 9*, 8–37. doi:10.1177/0891243905009001002

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender and Society, 1*, 125–151. doi:10.1177/0891243287001002002

Whatley, M. A. (2008). The dimensionality of the 15 item Attitudes Toward Women Scale. *Race, Gender, and Class, 15*, 265–273.

White, A. M., Strube, M. J., & Fisher, S. (1998). A Black feminist model of rape myth acceptance. *Psychology of Women Quarterly, 22*, 157–175. doi:10.1111/j.1471-6402.1998.tb00148.x

Wiseman, M. C., & Moradi, B. (2010). Body image and eating disorder symptoms in sexual minority men: A test and extension of objectification theory. *Journal of Counseling Psychology, 57*, 154–166. doi:10.1037/a0018937

Wong, M. R. (1984). MMPI Scale 5: Its meaning, or lack thereof. *Journal of Personality Assessment, 48*, 279–285. doi:10.1207/s15327752jpa4803_9

Worthington, R. L., Navarro, R. L., Savoy, H. B., & Hampton, D. (2008). Development, reliability and validity of the Measure of Sexual Identity Exploration and Commitment (MoSIEC). *Developmental Psychology, 44*, 22–33. doi:10.1037/0012-1649.44.1.22

Worthington, R. W., & Whittaker, T. A. (2006). Using exploratory and confirmatory factor analysis in scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806–838. doi:10.1177/0011000006288127

Yoder, J. D. (2007). *Women and gender: Making a difference*. Cornwall-on-Hudson, NY: Sloan.

Yoder, J. D., Perry, R. L., & Saal, E. I. (2007). What good is a feminist identity?: Women's feminist identification and role expectations for intimate and sexual relationships. *Sex Roles, 57*, 365–372. doi:10.1007/s11199-007-9269-2

CHAPTER 28

ASSESSING MEANING AND QUALITY OF LIFE

Michael F. Steger

What does it mean to live the "good life"? How can societies gauge how well they are serving their populace? What are the best ways to understand the effects of illness, disease, injury, and treatment on people? What matters most in life? These are questions that psychologists have increasingly turned their expertise toward answering. Although many of the variables psychologists study and assess touch on these complex questions, the constructs of *meaning in life* and *quality of life* are directly focused on helping individuals and societies gauge where they land on such questions, although they come from very different historical traditions.

Meaning in life came to the fore as a crucial way of understanding how people overcome life challenges and maximize their unique potentials within humanistic and existential branches of psychology. Drawing on centuries of preceding philosophical work, psychologists began to argue for the power—to ennoble or debilitate—of facing the responsibility to build our own existences in the face of inescapable death (e.g., Sartre, 1956). Within this tradition, the emphasis has been on that which essentially and truly makes life worth living. This perspective is adroitly captured by Nietzsche's maxim, which can be paraphrased as, "you can endure anything if you have a reason to live," as well as in the example of Frankl's (1963) survival of Nazi concentration camp horrors. At its heart, meaning in life is intended to serve as a way to capture whether people have a reason to live.

Quality of life, on the other hand, emerged as a key way for the field of medicine to consider important outcomes beyond people's disease status or death (Power, 2003). Although knowing whether a patient is dead or alive, or whether a nation's citizens suffer from higher rates of serious illness than other places, are factors of obvious importance, other outcomes are also important. For example, the treatment for many serious conditions is as serious as the disease. In the case of cancer, radiation and chemotherapy bring a host of side effects. Many psychopharmaceutical treatments also have side effects. For example, neuroleptic or antipsychotic medications, particularly older ones, carry a risk of extrapyramidal side effects, such as tardive dyskinesia, which can permanently impair patient functioning beyond any detriments associated with the psychological disorder meant to be treated (e.g., Advokat, Mayville, & Matson, 2000). Quality of life has been used to try to balance these factors and has helped shift thinking about disease and treatment away from a simple arithmetic of "diseased" versus "no disease" or "dead" versus "alive" toward a more holistic effort to ask what level and desirability of experience is likely to accompany disease or treatment. Drawing on the World Health Organization's (1948) definition of health, which includes both an absence of illness and a presence of wellness, quality-of-life research asks not simply how long might you live, but also asks how well you might live.

Research on both meaning in life and quality of life has expanded considerably in recent years. Quality of life research has exploded in recent decades, with exponential acceleration. A PsycINFO search for *quality of life* returned 13,903 titles since 2005

DOI: 10.1037/14048-028
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

alone (Figure 28.1, top panel; out of 20,891 between 2000 and 2010). A PsycINFO search of the terms *meaning in life* or *purpose in life* also reveal mounting research on meaning in life, although at a more modest level (Figure 28.1, bottom panel; 625 of the 1,009 titles returned between 2000 and 2010 were published since 2005).[1] Along with the increase in overall attention paid to meaning in life and quality of live has come a proliferation of measures. This is particularly the case for quality of life. As quality of life originally emerged to get a better idea of the ways in which disorder and treatment affect people's functioning, there was mounting recognition that many disorders and treatments were not directly comparable. For example, treating disease by removing the affected body part, as in amputation, makes a discrete and irreversible effect, whereas maintaining a healthy diet as part of managing diabetes creates a sustained effect that can fluctuate in intensity and raises the possibility of relapse. Without making light of any of the suffering and hardship that people experience, it seems fairly clear that if an instrument does not exist already to measure quality of life with regard to a specific illness, disability, disease, or disorder, it will be published soon. A quick reading of recent literature reveals quality of life measures intended for specific use among populations suffering from erosive esophagitis (Wyrich et al., 2010), cancer (e.g., Zebrack, 2009), and otitis media (Timmerman, Meesters, Speyer, & Anteunis, 2007) as well as those facing end-of-life issues (Henoch, Axelsson, & Bergman, 2010). In this psychometrics-intensive area, it is not uncommon to see articles describing the performance of a head-and-neck-specific quality of life instrument among asymptomatic patients with throat cancer in Greece (e.g., Nalbadian et al., 2010), late-stage dementia among long-term care facility residents in Spain, or cancer-related quality of life among young adult survivors of childhood cancer in the United States (Zebrack et al., 2010).

To a lesser extent, measure proliferation has taken root in meaning-in-life research as well. However, even in this smaller research area, there are over a dozen published measures. Although research

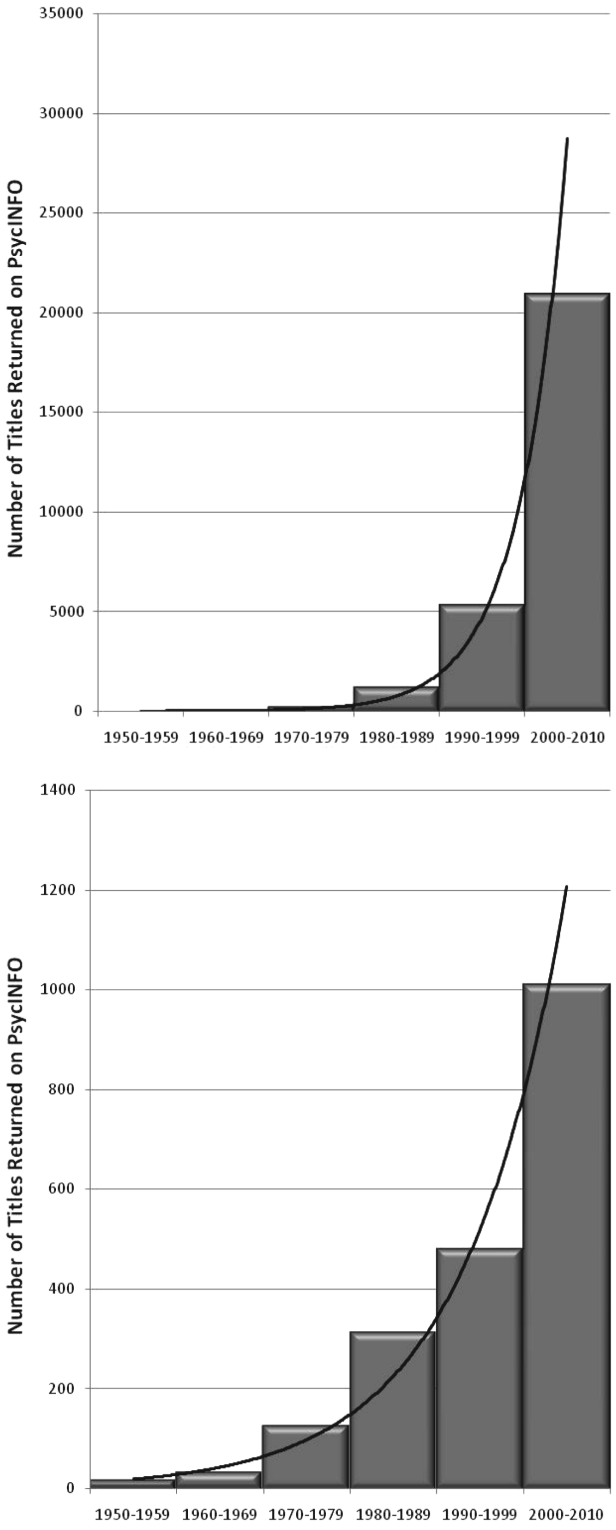


FIGURE 28.1. Increase in the numbers of indexed works returned in a PsycINFO search for the terms *quality of life* (top panel) and *meaning in life* or *purpose in life* (bottom panel).

[1]PsycINFO search conducted on June 14, 2010, using the terms *quality of life* and *meaning in life* or *purpose in life*.

on both meaning and quality of life has usually focused on broad, higher order appraisals of one's overall life, within these areas of research, periodic attention has been paid to how people weight particular domains. In meaning in life research, this branch of research has generated numerous attempts to measure people's sources of meaning in life (e.g., Schnell, 2009); in quality of life, it has spurred debates about how many domains of life there are that people use to make quality-of-life judgments and whether these domains contribute equally to people's judgments (e.g., Hagerty et al., 2001). Factors such as specific disease populations, specific domains, and specific demographic populations create a complex assessment landscape. As a result, there are far more individual assessments than could be reviewed in this chapter, Because of these factors, the focus of this chapter is on broad, general-purpose measures of meaning in life and quality of life. This chapter is not intended to be encyclopedic but rather simply reviews the psychometric evidence and suggest particularly useful or well-validated measures within each category.

MEANING IN LIFE

Meaning-in-life research really began in the mid-1960s, with the publication of the Purpose in Life test (PIL; Crumbaugh & Maholick, 1964). From this early time, ideas about meaning in life have been dominated by the writings of Viktor Frankl (1963) who identified several features of the meaningful life (e.g., creative acts and adopting resilient attitudes toward suffering), and decried what he perceived as a unique malady of the modern age—a bored, anxious malaise he called *noögenic neurosis*. From its empirical origins, meaning in life has been loosely defined, including qualities related to how people express themselves and their autonomy and how they adapt to adversity as well as how distressed they feel themselves to be. The meaningful life was viewed as a rich, interesting, personally expressive adventure, full of goals to pursue and passionate activity, and free from anxiety, boredom, and depression. Viewing meaning in life this way poses significant problems for both meaning-in-life theory and assessment. In particular, it appeared to take the field a while to resolve whether to include in the definition of meaning-in-life constructs such as creativity, activity level, energy, depression, anxiety, and boredom or whether to regard them as consequences of living either a meaningful or meaningless existence. Thus, the PIL includes items about excitement, goal-directed activity, and even suicidal ideation, and as a whole, meaning-in-life measurement trended toward inclusiveness of content. A similar lack of theoretical clarity may have led to the use of the Sense of Coherence Scale (Antonovsky, 1987) to measure meaning, despite the fact that its meaningfulness subscale pertained to whether people felt stressors were typically worth their coping efforts. Aside from criticism directed at the inclusion of inappropriate content in the PIL (e.g., Dyck, 1987), the practical effect is that the PIL is unsuitable for use in research targeting depression, schizophrenia, or other psychological disorders because of the items that assess excitement, energy, and suicide. Additionally, its factor structure has been extremely unstable across studies (e.g., Steger, 2006). Therefore, the PIL cannot be recommended. However, three other scales have been used fairly widely and have good or encouraging psychometric properties or provide features that may be advantageous for future meaning-in-life research and applied work (see Table 28.1). Thus, this chapter reviews the Life Regard Index (LRI; Battista & Almond, 1973), the Life Attitude Profile—Revised (LAP–R; Reker, 1992), and the Meaning in Life Questionnaire (MLQ; Steger, Frazier, Oishi, & Kaler, 2006).

The LRI

The LRI was developed to assess a theoretically driven idea of meaning in life, which argued that meaning in life consisted of having a cognitive framework for understanding one's experience in life and forming important goals, and an emotional sense of fulfillment on having established this framework (Battista & Almond, 1973). The LRI consists of two 14-item scales corresponding to the framework idea (e.g., about having particular goals that would bring personal satisfaction, about finding meaningful direction in living, and about becoming energized by strong interest in particular activities)

TABLE 28.1

Table of Frequently Used and Psychometrically Sound Measures of Meaning in Life

Name	No. of items	Subscales	Strengths	Weaknesses
Life Regard Index (Battista & Almond, 1973)	28	Framework	Tied to a theory	Only two dimensions assessed
		Fulfillment	Relatively large research literature	Fulfillment dimension conflates meaning measurement with well-being and other content areas
			Reliable subscale scores	Many items per subscale
				Poor structural validity
				Lack of empirical scale construction techniques
Life Attitude Profile—Revised (Reker, 1992)	48	Purpose	Multidimensional	Full scale is longer than most other meaning in life measures
		Coherence	Subscales are relatively brief	Possible overinclusion of content in meaning in life scales (e.g., exciting life, restlessness)
		Choice/ Responsibleness	Specifies content of meaning in life	Several subscales have not received substantial research attention
		Death Acceptance	Psychometric evidence continues to be published	Failure to support purpose-coherence distinction
		Existential Vacuum	Use of empirical scale construction techniques	
		Goal Seeking		
Meaning in Life Questionnaire (Steger, Frazier, Oishi, & Kaler, 2006)	10	Presence of Meaning in Life	Short	Only two dimensions assessed
		Search for Meaning in Life	Easy to use	Combines *meaning* and *purpose* items
			Assesses two key dimensions with different correlates	Items are subjective, and thus relatively contentfree (e.g., items focus on *meaning* and *purpose*)
			Use of empirical scale construction techniques	
			Robust psychometric properties	
			Psychometric evidence continues to be published	
			Face valid	
			Free of extraneous content	
			Available in many languages	

and the fulfillment idea (e.g., about feeling deep fulfillment in life, about feeling energetic and excited, and reverse-scored items about a lack of outstanding things happening in life). Similar to the PIL, researchers have questioned the inclusion of extraneous content, particularly in the Fulfillment subscale, which asks whether people "feel really good" of "have real passion" about their lives (e.g., Steger, 2007). There is little to suggest that feeling good about one's life is something specific to meaning in life. Also, factor analyses have been somewhat inconsistent in supporting the LRI's factor structure (e.g., de Klerk, Boshoff, & van Wyk, 2009; Steger, 2007). As a response, researchers should only use the Framework subscale, which is typically highly reliable across samples and shows convergent validity as one would expect (e.g., Debats, van der Lubbe, & Wezeman, 1993; Steger et al., 2006).

Therefore, despite some weaknesses, due to its theoretical grounding, fairly common use and typically reliable scores, the Framework subscale of the LRI appears worth considering for research on meaning in life. However, research should focus on identifying a core set of items from this subscale that can provide structural stability and theoretical fidelity.

The LAP–R

The LAP–R, and its early incarnation, the Life Attitude Profile (Reker & Peacock, 1981), drew content areas from classic logotherapy scholarship, including Frankl's (1963) writings. It provides the only truly multidimensional, theoretically driven assessment of meaning in life that also has some evidence of validity and reliability. The LAP–R uses 48 items to assess six dimensions: purpose (e.g., about having a mission and a sense of direction), coherence (e.g., about seeing the meaning of life around one), choice/responsibleness (e.g., about viewing self-direction as very important), death acceptance (e.g., about feeling relatively unconcerned about death), existential vacuum (e.g., about feeling one has a destiny one cannot identify), and goal seeking (e.g., liking new and different things). Some subscales can be combined to produce additional index scores. The Personal Meaning Index is created through combining the Purpose and Coherence subscales, and the Existential Transcendence index is created by summing the Purpose, Coherence, Choice/Responsibleness, and Death Acceptance subscales and subtracting both the Existential Vacuum and Goal-Seeking subscales. Research from both the LAP–R and its original shows generally acceptable reliabilities for subscale scores and convergent validity as one would expect (e.g., Reker, Peacock, & Wong, 1987). From a structural standpoint, there have been difficulties establishing that purpose and coherence are distinct factors, although the Personal Meaning Index, which combines the two, shows unifactorial invariance across age groups (Reker, 2005). Nonetheless, additional research is necessary to verify the structural composition of the LAP–R as a whole as well as the usefulness of each of its subscales.

Thus, for those interested in pursuing some of the dimensions suggested by logotherapy, the LAP–R is the best measurement choice. In addition, the 16-item Personal Meaning Index appears to be a psychometrically sound choice for assessing the presence or absence of meaning in life, although it is longer than the Framework subscale of the LRI and includes some questionable item content (e.g., about life having exciting good things in it, and about feeling fulfilled in achieving life goals).

The MLQ

The MLQ was developed to address perceived shortcomings in the assessment of meaning in life (Steger et al., 2006). In particular, it was developed to be brief, structurally robust, psychometrically sound, and free of extraneous content. The MLQ consists of 2 five-item subscales: Presence of Meaning in Life (e.g., about one's life having a clear sense of purpose, and about understanding one's life meaning) and Search for Meaning in Life (e.g., about searching for meaning in life, and about looking for something that makes life feel meaningful). Research has established that MLQ scores are reliable and reasonably stable over 1 year's time (Steger & Kashdan, 2007), and the two-factor structure has been replicated in samples from several countries (e.g., Spain [Steger, Frazier, & Zacchanini, 2008], Japan [Steger, Kawabata, Shimai, & Otake, 2008]) and across age groups (Steger, Oishi, & Kashdan, 2009). In

addition, the Presence of Meaning and Search for Meaning subscales have been shown to have different correlates (Dunn & O'Brien, 2009; Steger, Kashdan, Sullivan, & Lorentz, 2008; Vess, Arndt, Cox, Routledge, & Goldenberg, 2009). Because the MLQ is brief and shows evidence of face validity, with short, clear items, it is being used for multinational health and well-being surveillance research (e.g., Samman, 2007; http://wellbeingstudy.com). Some information about the MLQ is available on the Internet, a downloadable version is freely available (http://michaelfsteger.com/MLQ.aspx), and translated versions in more than two dozen languages are also available. However, the Web-based resources that exist for the quality-of-life measures (discussed in the next section) do not generally exist for measures of meaning in life.

Thus, there is a great deal of research evidence to support the use of the MLQ. However, it only assesses two dimensions, and because of this the MLQ may be more of an efficient means for assessing meaning in life, rather than a tool to be used for elaborating meaning-in-life theory. At the same time, it is the only instrument that reliably measures people's search for meaning in life, which is an historically important element of meaning-in-life theories, denoting the human "will to meaning" (Frankl, 1963).

Other Issues in Assessing Meaning in Life

Among the critical challenges facing assessment of meaning in life is the necessity of demonstrating incremental validity beyond many other well-being variables. To date, only the MLQ has been shown to be related yet distinct from the key well-being variables of self-esteem, life satisfaction, and optimism using the gold standard for establishing convergent and discriminant validity, the multitrait–multimethod matrix. Future research should continue this effort so that unique contributions of meaning in life to health and other valued outcomes can be established.

It is also relatively unknown whether it matters from where people draw meaning in their lives. Thus, at least theoretically, it is somewhat possible that someone who finds meaning through swindling people out of their money could look as well adjusted as someone who finds meaning through administering hospice care to dying patients. Some new assessment approaches are attempting to tackle this issue (e.g., Schnell, 2009), but research that goes beyond simply identifying from where people say they draw meaning in life is practically nonexistent.

QUALITY OF LIFE

Unlike the meaning-in-life research area, which has been driven largely by efforts to define the construct, quality of life research seems to have been concerned to a greater degree by an interest in determining the number and specificity of domains of functioning. The overall effort is directed toward capturing key elements of how people view their situation. In fact, one comprehensive review of quality-of-life measures included *Money* magazine's "Best Places to Live" list (Hagerty et al., 2001). Although this example was drawn from a project intended to evaluate assessments that could be used to inform national policy, it speaks to a highly pragmatic approach common in quality of life research.

Befitting its origins, definitions of quality of life usually carry connotations of subjective judgments of health, and instruments are most frequently used in medical and public health contexts. It is this author's personal impression that the extensive clinical applications of this area of research has led to a rather large number of fee-for-use measures, akin to the assessment of psychological disorders. Although it is certainly reasonable to commercialize assessments, it may be that one influence of this commercial approach to assessment is that many of the quality-of-life measures that have been developed do not have extensive validation research supporting their use (apart from what is presented in the corresponding manuals). Thus, there are many dozens of quality-of-life measures, yet relatively few of them have generated a strong research literature. In this section, general-purpose quality-of-life assessments that have broad potential for application will be considered, with selections and recommendations based on the available research (see Table 28.2), which does not necessarily preclude the inclusion of fee-for-use measures. The most widely used and researched quality-of-life assessment instruments

TABLE 28.2

Table of Frequently Used and Psychometrically Sound Measures of Quality of Life

Name	No. of items	Subscales/dimensions	Strengths	Weaknesses
EuroQOL (EuroQOL Group, 1990)	5	Mobility	Short	May neglect some important dimensions of quality of life
		Self-Care	Easy to use	No positive anchor to item rating scale
		Ability to Perform Usual Activities	Easy for respondents to understand	Total quality of life score poses interpretation problems
		Pain and Discomfort	Available in several forms, varying generally by length	
		Anxiety and Depression	Available in many languages	
			Large body of research	
Medical Outcomes Survey Short Form 36 (Ware & Sherbourne, 1992)	36	Physical Functioning	Multidimensional coverage	Uneven coverage of dimensions
		Role Performance: Physical Impairment	Easy to use	Potential item formatting problems (e.g., *getting sick* and *getting worse* items may form own factor)
		Bodily Pain	Available in shorter forms	Items from several scales focus on general impairment, and many vitality items are also symptoms of mental health
		General Health	Available in many languages	May neglect some important dimensions of quality of life health disorders
		Mental Health	Large body of research	
		Vitality		
		Role Performance: Emotional Impairment		
		Social Functioning		

appear to be the EuroQOL EQ-5D (EuroQOL Group, 1990), the Medical Outcomes Survey Short Form 36 (SF-36; Ware & Sherbourne, 1992), and a set of quality-of-life scales developed by the World Health Organization. Each of these measures is multidimensional but with a slightly different array of dimensions assessed. Reviewing these measures is difficult because of stringent copyright and intellectual property protection statements. For example, referring to sample items and even rating scale anchors is disallowed, and one widely used measure restricts even citations and inclusion in bibliographies. More information is available for these measures than can be provided in this context because these policies restrict what can be presented (and even if they can be presented at all), so readers are encouraged to find out more on their own if they want to uses these three measures.

The EQ-5D

The shortest quality-of-life measure reviewed in this chapter is the five-item EQ-5D. This scale consists of

five dimensions. The dimensions cover mobility, self-care, ability to perform usual activities, pain and discomfort, and anxiety and depression. Each dimension is assessed with three "levels" of statements that are anchored to self-rated judgments of how difficult functioning in each dimension is. Total scores are expressed as a five-digit "health state" value in which each dimension receives a score from 1 to 3 in its own column. For example, a health state of 11111 would correspond to the most ideal score attainable on the EQ-5D, indicating no quality-of-life problems. A health state of 11131 would indicate the presence of severe pain or discomfort with no other quality-of-life problems, and so forth. An overall health scale can also be used, with respondents indicating their health on a rating scale ranging from 0 to 100.

There is a large research literature supporting the use of the EQ-5D (e.g., Shaw, Johnson, & Coons, 2005), and it is the briefest of the quality of life measures reviewed here. However, it is not without some complications in its application. First, as it only consists of five items, there are concerns about content that is not captured by the EQ-5D, such as spirituality. Second, some of the items combine distinct constructs, such as anxiety and depression, which obscures what respondents are reporting on items like these. Third, the use of health states, rather than quality of life scores, may not be intuitive for all users.

There appears to have been some effort to overcome the ceiling effects that have been noted for the EQ-5D, with a five-level version also available for use. In addition, a youth version has been developed. There is a very helpful website that supports the EQ-5D (http://www.euroqol.org). Fees for using the EQ-5D are determined according to the use for which the measure is intended.[2]

The Medical Outcomes Survey SF-36

The SF-36 was developed to assess multiple dimensions of quality of life in a brief format. The 36 items cover four physical dimensions and four mental dimensions. The physical dimensions are physical functioning, physical impairment in important roles, bodily pain, and general health. The psychological dimensions are vitality, social functioning, emotional impairment in important roles, and mental health. The SF-36 has been used in thousands of studies, as has the revision of the SF-36 (SF-36v2). The SF-36v2 standardized the number of response options (the SF-36 used different response anchors for each item) and attempted to clarify wording, but it is otherwise very similar to the SF-36.

The SF-36 is probably the most widely used quality-of-life measure; therefore, there is an enormous amount of information connecting SF-36 scores to other measures of functioning, specific disease symptoms, and uses in populations of individuals experiencing specific diseases of interest. The psychometric properties appear adequate for the two major component scores (physical and mental, as described earlier) and most of the specific dimensions (Ware, Kosinski, & Keller, 1994), and factor analyses seem to support many of the structural qualities of the SF-36, although some factor analyses suggest that there are nine rather than eight factors (for a review, see de Vet, Adèr, Terwee, & Pouwer, 2005). Thus, it seems well recommended for use. At the same time, there are limitations to the SF-36. The SF-36 has been criticized for combining both objective items, which capture behavioral and functional limitations (e.g., physical impairment in important life roles), and subjective items (e.g., vitality; Power, 2003). However, this seems largely to reflect a difference in the physical versus mental domains. A different matter is that the dimensions are assessed using widely varying numbers and kinds of items. For example, the bodily pain and social functioning items are assessed using two items, both of which ask about two different aspects of the same construct (magnitude and degree of interference for pain; degree of interference and time lost for social functioning). On the other hand, the Physical Functioning subscale is assessed by asking about degree of impairment in 10 different activities. This disparity should lead to some caution when the primary objective of using a quality-of-life measure is identifying specific areas of improvement or deterioration. Finally, it seems possible that the high

[2]Interested parties must obtain specific information about license fees by completing and submitting registration forms on the measurement websites.

number of items that ask about "impairment" of some form on the Physical Functioning, Physical Impairment, Pain, and Social Functioning subscales could artificially increase score concordance. Such items could also present some respondents with difficulty in judging whether their impairment is due to pain, physical health problems, or emotional problems. Similarly, many of the vitality items express symptoms of psychological disorders that are also assessed by the Mental Health subscale (e.g., lack of energy is a symptom of major depressive disorder).

There are several variations of the SF-36, such as the aforementioned SF-36v2, a 12-item version, and an eight-item version. The SF-36 also has a website that provides many resources to interested parties (http://www.sf-36.org), including scoring services. As with the EQ-5D, fees and licensing restrictions are determined by the publisher depending on proposed use.[2]

Other Issues in Assessing Quality of Life

One critique that can be made of all of the aforementioned quality-of-life measures is that they neglect dimensions that some might find useful or even critical. Hagerty and colleagues (2001) reviewed a large number of quality-of-life measures and identified several domains that they recommended should be measured. One measure appears to address each of these domains, the Comprehensive Quality of Life Scale (Cummins, 1999; Cummins, McCabe, Romeo, & Gullone, 1994). Although research using this measure is scarce, it may be a compelling choice for those interested in capturing a full range of functioning across life domains.

However, if more specific measures of quality of life are needed, other reviews have been published. For example, Jacobsen, David, and Cella (2002) reviewed issues related to measuring quality of life in clinical trials for cancer treatment, including a description of leading cancer-related quality-of-life measures. Rajmil et al. (2004) reviewed the content included in health-related quality-of-life measures for use among children and adolescents.

A more important critique of quality-of-life measures concerns the aggressively proprietary nature of instrumentation in this field. Almost all meaning-in-life measures are freely available for educational, research, and clinical applications in which no fee is to be charged for access to the measures. This is starkly different from quality-of-life measurement. This chapter's author cannot recommend any of these widely used measures because of aggressive, stringent intellectual property issues. Instead, the field desperately needs a freely available, multiple-domain quality-of-life measure to help consolidate an enormously fragmented and unwieldy literature. Until that time, researchers should consider using instruments that are freely available for research, even if that means they must use one of the many condition-specific instruments out there. In a field dominated by for-profit instruments, only well-funded research programs are likely to find these instruments useful, and the field likely will stay mired in a repetitious litany of studies reporting that people with diseases that affect their health report lower quality of life. Currently, there are few options for advancing the field to a better understanding of what factors best promote quality of life. On the basis of existing, multidimensional measures, the Comprehensive Quality of Life Scale, developed by Cummins appears to be the best choice. More information on this scale is available at the website for the Australian Centre on Quality of Life (http://www.deakin.edu.au/research/acqol/instruments/comqol-scale).

CONCLUSION

As research on meaning in life and quality of life continues to accelerate, assessment becomes a more critical issue. Particularly in the case of quality of life, data from research are being sought for guidance with regard to national policy, the clinical effectiveness of medical treatments, and other high-stakes decisions. The demands placed on assessment instruments will also continue to mount. Researchers should strive to use the most psychometrically sound measures that can assess the domains of their interest according to their needs. The assessments reviewed in this chapter seem to show the best combination of attributes. At the same time, issues persist in both areas of research regarding definitions of the constructs and the number of dimensions necessary to assess it, and research should be encouraged that helps settle these questions.

References

Advokat, C. D., Mayville, E. A., & Matson, J. L. (2000). Side effect profiles of atypical antipsychotics, typical antipsychotics, or no psychotropic medications in persons with mental retardation. *Research in Developmental Disabilities, 21*, 75–84. doi:10.1016/S0891-4222(99)00031-1

Antonovsky, A. (1987). *Unraveling the mystery of health: How people manage stress and stay well.* San Francisco, CA: Jossey-Bass.

Battista, J., & Almond, R. (1973). The development of meaning in life. *Psychiatry, 36*, 409–427.

Crumbaugh, J. C., & Maholick, L. T. (1964). An experimental study in existentialism: The psychometric approach to Frankl's concept of noogenic neurosis. *Journal of Clinical Psychology, 20*, 200–207. doi:10.1002/1097-4679(196404)20:2<200::AID-JCLP2270200203>3.0.CO;2-U

Cummins, R. A. (1999). A psychometric evaluation of the comprehensive Quality of Life Scale. In L. L. Yuan, B. Yuen, & C. Low (Eds.), *Urban quality of life: Critical issues and options* (5th ed., pp. 20–33). Singapore: Singapore University Press.

Cummins, R. A., McCabe, M. P., Romeo, Y., & Gullone, E. (1994). The Comprehensive Quality of Life Scale (ComQol): Instrument development and psychometric evaluation on college staff and students. *Educational and Psychological Measurement, 54*, 372–382. doi:10.1177/0013164494054002011

Debats, D. L., van der Lubbe, P. M., & Wezeman, F. R. A. (1993). On the psychometric properties of the Life Regard Index (LRI): A measure of meaningful life. *Personality and Individual Differences, 14*, 337–345. doi:10.1016/0191-8869(93)90132-M

de Klerk, J. J., Boshoff, A. B., & Van Wyk, R. (2009). Measuring meaning in life in South Africa: Validation of an instrument developed in the USA. *South African Journal of Psychology, 39*, 314–325.

de Vet, H. C. W., Adèr, H. J., Terwee, C. B., & Pouwer, F. (2005). Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Quality of Life Research, 14*, 1203–1218. doi:10.1007/s11136-004-5742-3

Dunn, M., & O'Brien, K. (2009). Psychological health and meaning in life: Stress, social support, and religious coping in Latina/Latino immigrants. *Hispanic Journal of Behavioral Sciences, 31*, 204–227. doi:10.1177/0739986309334799

Dyck, M. J. (1987). Assessing logotherapeutic constructs: Conceptual and psychometric status of the Purpose in Life and Seeking of Noetic Goals Tests. *Clinical Psychology Review, 7*, 439–447. doi:10.1016/0272-7358(87)90021-3

EuroQOL Group. (1990). EuroQOL—A new facility for the measurement of health-related quality of life. *Health Policy (Amsterdam)*, 16, 199–208. doi:10.1016/0168-8510(90)90421-9

Frankl, V. E. (1963). *Man's search for meaning: An introduction to logotherapy*. New York, NY: Washington Square Press.

Garre-Olmo, J., Planas-Pujol, X., Lopez-Pousa, S., Weiner, M. F., Turon-Estrada, A., Juvinyà, D., . . . Vilalta-Franch, J. (2010). Cross-cultural adaptation and psychometric validation of a Spanish version of the Quality of Life in Late-Stage Dementia Scale. *Quality of Life Research, 19*, 445–453. doi:10.1007/s11136-010-9594-8

Hagerty, M. R., Cummins, R. A., Ferriss, A. L., Land, K., Michalos, A. C., Peterson, M., . . . Vogel, J. (2001). Quality of life indexes for national policy: Review and agenda. *Social Indicators Research*, *55*, 1–96. doi:10.1023/A:1010811312332

Henoch, I., Axelsson, B., & Bergman, B. (2010). The Assessment of Quality of life at the End of Life (AQEL) questionnaire: A brief but comprehensive instrument for use in patients with cancer in palliative care. *Quality of Life Research, 19*, 739–750. doi:10.1007/s11136-010-9623-7

Jacobsen, P. B., Davis, K., & Cella, D. (2002). Assessing quality of life in research and clinical practice. *Oncology*, *16*(Suppl. 10), 133–139.

Nalbadian, M., Nikolaidis, V., Nikolaou, A., Themelis, C., Kouloulas, A., & Vital, V. (2010). Psychometric properties of the EORTC head and neck-specific quality of life questionnaire in disease-free Greek patients with cancer of pharynx and larynx. *Quality of Life Research, 19*, 761–768. doi:10.1007/s11136-010-9628-2

Power, M. J. (2003). Quality of life. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 427–441). Washington, DC: American Psychological Association. doi:10.1037/10612-027

Rajmil, L., Herdman, M., Fernandez de Sanmamed, M.-J., Detmar, S., Bruil, J., Ravens-Sieberer, U., . . . Auquier, P. (2004). Generic health-related quality of life instruments in children and adolescents: A qualitative analysis of content. *Journal of Adolescent Health*, *34*, 37–45. doi:10.1016/S1054-139X(03)00249-0

Reker, G. T. (1992). *Life Attitude Profile—Revised.* Peterborough, Ontario, Canada: Student Psychologists Press.

Reker, G. T. (2005). Meaning in life of young, middle-aged, and older adults: Factorial validity, age, and gender invariance of the Personal Meaning Index (PMI). *Personality and Individual Differences, 38*, 71–85. doi:10.1016/j.paid.2004.03.010

Reker, G. T., & Peacock, E. J. (1981). The Life Attitude Profile (LAP): A multidimensional instrument for assessing attitudes toward life. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 13*, 264–273. doi:10.1037/h0081178

Reker, G. T., Peacock, E. J., & Wong, P. T. P. (1987). Meaning and purpose in life and well-being: A life-span perspective. *Journal of Gerontology, 42*, 44–49.

Samman, E. (2007), Psychological and subjective well-being: A proposal for internationally comparable indicators. *Oxford Development Studies 35*, 459–486. doi:10.1080/13600810701701939

Sartre, J. P. (1956). *Being and nothingness* (H. Barnes, Trans.). New York, NY: Philosophical Library.

Schnell, T. (2009). The Sources of Meaning and Meaning in Life Questionnaire (SoMe): Relations to demographics and well-being. *Journal of Positive Psychology, 4*, 483–499. doi:10.1080/17439760903271074

Shaw, J. W., Johnson, J. A., & Coons, S. J. (2005). US Valuation of the EQ-5D Health States: Development and testing of the D1 valuation model. *Medical Care, 43*, 203–220. doi:10.1097/00005650-200503000-00003

Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment, 86*, 263–272. doi:10.1207/s15327752jpa8603_03

Steger, M. F. (2007). Structural validity of the Life Regards Index. *Measurement and Evaluation in Counseling and Development, 40*, 97–109.

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The Meaning in Life Questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology, 53*, 80–93. doi:10.1037/0022-0167.53.1.80

Steger, M. F., Frazier, P., & Zacchanini, J. L. (2008). Terrorism in two cultures: Traumatization and existential protective factors following the September 11th attacks and the Madrid train bombings. *Journal of Trauma and Loss, 13*, 511–527. doi:10.1080/15325020802173660

Steger, M. F., & Kashdan, T. B. (2007). Stability and specificity of meaning in life and life satisfaction over one year. *Journal of Happiness Studies, 8*, 161–179. doi:10.1007/s10902-006-9011-8

Steger, M. F., Kashdan, T. B., Sullivan, B. A., & Lorentz, D. (2008). Understanding the search for meaning in life: Personality, cognitive style, and the dynamic between seeking and experiencing meaning. *Journal of Personality, 76*, 199–228. doi:10.1111/j.1467-6494.2007.00484.x

Steger, M. F., Kawabata, Y., Shimai, S., & Otake, K. (2008). The meaningful life in Japan and the United States: Levels and correlates of meaning in life. *Journal of Research in Personality, 42*, 660–678. doi:10.1016/j.jrp.2007.09.003

Steger, M. F., Oishi, S., & Kashdan, T. B. (2009). Meaning in life across the life span: Levels and correlates of meaning in life from emerging adulthood to older adulthood. *Journal of Positive Psychology, 4*, 43–52. doi:10.1080/17439760802303127

Timmerman, A. A., Meesters, C. M. G., Speyer, R., & Anteunis, L. J. C. (2007). Psychometric qualities of questionnaires for the assessment of otitis media impact. *Clinical Otolaryngology, 32*, 429–439. doi:10.1111/j.1749-4486.2007.01570.x

Vess, M., Arndt, J., Cox, C., Routledge, C., & Goldenberg, J. (2009). Exploring the existential function of religion: The effect of religious fundamentalism and mortality salience on faith-based medical refusals. *Journal of Personality and Social Psychology, 97*, 334–350. doi:10.1037/a0015545

Ware, J. E., Kosinski, M., & Keller, S. K. (1994). *SF-36® physical and mental health summary scales: A user's manual*. Boston, MA: The Health Institute.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health status survey (SF-36): 1. Conceptual framework and item selection. *Medical Care, 30*, 473–483. doi:10.1097/00005650-199206000-00002

World Health Organization. (1948). *Basic documents* (45th ed.). Geneva, Switzerland: Author.

Wyrwich, K. W., Mody, R., Larsen, L. M., Lee, M., Harnam, N., & Revicki, D. A. (2010). Validation of the PAGI-SYM and PAGI-QOL among healing and maintenance of erosive esophagitis clinical trial participants. *Quality of Life Research, 19*, 551–564. doi:10.1007/s11136-010-9620-x

Zebrack, B. (2009). Developing a new instrument to assess the impact of cancer in young adult survivors of childhood cancer. *Journal of Cancer Survivorship: Research and Practice, 3*, 174–180. doi:10.1007/s11764-009-0087-0

Zebrack, B. J., Donohue, J. E., Gurney, J. G., Chesler, M. A., Bhatia, S., & Landier, Q. (2010). Psychometric evaluation of the Impact of Cancer (IOC-CS) Scale for young adult survivors of childhood cancer. *Quality of Life Research, 19*, 207–218. doi:10.1007/s11136-009-9576-x

CHAPTER 29

ASSESSMENT IN REHABILITATION PSYCHOLOGY

Jennifer E. Stevenson, Kathleen B. Kortte, Cynthia F. Salorio, and Daniel E. Rohe

Rehabilitation psychology is a specialty area of practice that "focuses on the study and application of psychological knowledge and skills on behalf of individuals with disabilities and chronic health conditions in order to maximize health and welfare, independence and choice, functional abilities, and social role participation" (Frank & Caplan, 2010, p. 3). Rehabilitation psychology has been in existence for over 50 years, developing into a recognized, board-certified area of practice in 1997. Rehabilitation psychologists practice clinically in a number of settings, including, although not limited to, subacute, acute, and postacute medical rehabilitation settings, private practice, vocational rehabilitation settings, school settings, and military and veterans' hospitals. This chapter focuses on the most common setting for practice, which is the medical rehabilitation setting.

Assessment in rehabilitation psychology practice provides key information about biological, psychological, and social factors that may facilitate or hinder the rehabilitation and recovery process within the context of complex medical conditions and traumatic injury. Assessment facilitates the process by guiding intervention and discharge planning through the identification of potential physical, cognitive, and emotional barriers. Psychologists practicing within a medical rehabilitation setting must have a thorough understanding not only of psychiatric, emotional, and behavioral disorders but also of a variety of medical diagnoses, their associated prognoses, and contemporary treatments. To enhance the rehabilitation process, it is imperative not only to identify barriers to convalescence but also to discover intrapersonal strengths and resiliency factors that foster continued recovery and social reintegration.

A GUIDING FRAMEWORK

Regardless of the rehabilitation setting, population, or age group, there are three key components of the individual that form the basis for rehabilitation psychology assessments: physical functioning, psychological functioning, and cognitive functioning. This "Guiding Framework" provides a structure to build a comprehensive and holistic evaluation of the individual. These component areas are based on the biopsychosocial model, a model that highlights the multidimensional nature of the individual as well as individual characteristics that can either hinder or facilitate functioning in everyday life (Dunn & Dougherty, 2005). Evaluations of an individual must take into consideration the environment in which they reside. The role the environment plays in rehabilitation outcomes is pivotal, with the person–environment interaction being a crucial aspect in determining whether an individual is disabled or not (Dunn & Dougherty, 2005; Lewin, 1939).

The Guiding Framework is outlined in Table 29.1 and is also reflected in other tables within this chapter. Although aspects of it may appear simplistic on the surface, the Guiding Framework provides a reminder of the interplay between the multiple aspects of human functioning within the context of the environment. At the time of evaluation, the individual's environment might be the acute inpatient

DOI: 10.1037/14048-029
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

TABLE 29.1

A Guiding Framework for Rehabilitation Psychology Assessments

Component area	Focus of assessment
Person	
Cognitive functioning	Capacity
	Mental status
	Neuropsychological functioning
Physical functioning	Fatigue
	Health management behavior
	Pain
	Sleep
Psychological functioning	Adaptive functioning
	Emotional adjustment
	Interpersonal/social functioning
	Personality
	Psychiatric conditions
Environment	
Cultural context	Cultural environment
Physical context	Community environment
	Living environment
Role context	Home environment
	Rehabilitation environment
	School environment
	Vocational environment
Social context	Social environment

rehabilitation unit, which offers its own unique social, physical, and cultural environmental factors, including the physical make-up of the unit, whether the individual has to share a room with another patient, the structure of daily unit activities, and the dynamics of the rehabilitation team. This may be contrasted with the outpatient, or postacute, rehabilitation environment in which environmental factors may include living in a home with limited accessibility for wheelchair use or living with family members with whom the person has not lived for years. The interplay between the person and environmental factors presents a myriad of possible life issues and life successes. The environment that facilitates success for one person can lead to difficulties for another because of the within-person factors. Thus, consideration must be given to both the person and the environment.

THE ASSESSMENT PROCESS

Rehabilitation psychologists often are integrated members of an interdisciplinary rehabilitation team. The team structure lends itself to a somewhat different *referral* process than in typical mental health settings, although the *assessment* process is structured similarly. In collaboration with the other members of the team, rehabilitation psychologists determine the need for, and type of, assessment for individual patients. This determination is based on many factors including cognitive, behavioral, personality, mood, or pain problems that may affect the rehabilitation process. Identification of the referral question is an important initial step so that the evaluation can be properly focused. This critical step facilitates the appropriate design of the assessment and also ensures that unnecessary and/or misguided assessment procedures do not occur.

Content of the evaluation is influenced by many factors including the medical diagnosis, associated factors (e.g., suspected cognitive impairment, psychiatric/emotional disturbance, engagement in behaviors that may be detrimental to health maintenance, and questions about decision-making capacity), and the referral question. Assessing biological/physical, psychological/emotional, and social/environmental factors that are contributing to rehabilitation and recovery is an ongoing process that begins during the acute phase of rehabilitation and ought to continue through the subacute and outpatient phases of rehabilitation. Early assessment of barriers to the rehabilitation process is important to enable the patient to derive the maximum benefit from rehabilitation. There is also burgeoning evidence that there are person factors, such as hopefulness and positive affect, that when constructively identified, can facilitate the rehabilitation process and outcomes (Kortte, Gilbert, Gorman, & Wegener, 2010; Man et al., 2004; Vellone, Rega, Galletti, & Cohen, 2006). Assessment of these person factors may offer the clinician information about the individual that could offer potential avenues for supporting and bolstering that individual.

In the rehabilitation setting, typical referral questions include emotional adjustment issues, somatic symptoms that can be behaviorally managed, or cognitive decline. The overarching goal is to identify the barriers to the initial rehabilitation process and to social and vocational engagement. Assessment of adaptive functioning and positive psychological

skills can help to guide interventions that enhance life functioning. To accomplish this goal, the assessment process should include a multimethod approach to maximize the collection of relevant information, including a medical records review; a clinical interview; and the administration of objective measures designed to assess psychological, cognitive, behavioral, and vocational functioning.

Medical Records Review

In a medical rehabilitation setting, assessment begins with a review of the patient's medical record. Gathering background data such as a patient's past medical history including previous injuries, hospitalizations, and use of medications can provide key information that can guide not only the assessment but also the intervention and recommendations for a variety of services throughout recovery. An exhaustive review of an individual's past medical history may not always be feasible or necessary. Vital aspects of the medical chart review typically include review of medical diagnostic tests (e.g., computed tomography [CT] scans, magnetic resonance imaging [MRI] scans, X-rays, and laboratory test results including toxicology screening); past injuries and hospitalizations, especially those that involved brain injury/loss of consciousness, central nervous system impact, orthopedic injury, or other types of injury that may have resulted in acute and chronic pain; and information pertaining to previously prescribed medications including whether the individual always took the medication as prescribed or if there are concerns pertaining to abuse/dependence (e.g., as with opioid pain medications). An evaluation of current psychiatric and medical diagnoses also may include information gathered through the clinical interview; collateral information gleaned from family members, significant others, and friends as well as information gathered from the patient's medical record, and ought to include only information pertinent to current areas of distress and the referral question.

Clinical Interview

The clinical interview in the medical rehabilitation setting does not differ substantially from one performed in community, psychiatric, and counseling settings (see Chapter 7, this volume). The major difference is the focus on medical factors that could be contributing to the individual's psychological presentation and the effects that these factors may have on engagement in the rehabilitation process. The interview incorporates behavioral observations and a line of questioning to help glean information about current diagnoses (psychiatric and medical), the patient's and family's understanding of the reason for referral, and relevant aspects of the biopsychosocial history and current level of function. On the basis of work with the biopsychosocial model, gathering information pertaining to medical, mental health, and substance use history; social history (e.g., living situation, perceived level of and type of social support); educational history (i.e., including history of learning disorder or other relevant diagnoses); vocational history; avocational interests; and interactions with the legal system are all key parts of the clinical interview.

A determination of the patient's current functional status, as well as status before the onset of the current medical condition, assists a rehabilitation team in understanding the roles, values, and responsibilities of the individual and the effects that functional changes may have on daily life participation. Additionally, determining the patient's and family members' goals for rehabilitation and recovery can assist the psychologist in understanding the perspectives, values, and beliefs that they hold regarding functioning and disability overall. Finally, information about the physical, social, role, and cultural environments helps to formulate the context within which the individual person is expected to function. All of these can help to inform current treatment as well as relevant recommendations for tailoring the rehabilitation and community re-entry process.

During the clinical interview, rehabilitation psychologists are particularly interested in the assessment of client strengths and assets as key elements in determining personal resiliency factors that may facilitate the rehabilitation process. A strengths-based assessment approach incorporates information about an individual's personality characteristics, their typical means of coping with stressful situations, and their current psychological/emotional

response to their medical condition and disability status. Although some rehabilitation psychologists use assessment measures in clinical settings such as the Connor–Davidson Resilience Scale (Connor & Davidson, 2003) and the COPE (Carver, Scheier, & Weintraub, 1989), these instruments mostly have been used in research on the effects of positive psychological variables on rehabilitation.

In addition to the patient's verbal responses provided during the interview, important information can be gleaned simply by observing aspects of an individual's behavior during the interview. Behavioral observations can be made regarding the patient's physical, emotional, and cognitive status to gain a better understanding of the individual's functioning (see Table 29.2). Important indicators of impairment can be identified through careful observation of behaviors that begin with the initial patient contact and those that occur throughout the assessment process. Changes in behavior as the assessment progresses may help to identify physical, emotional, and cognitive difficulties that may wax and wane because of differential levels of arousal, pain intensity, confusional states, or reactions to aspects of the assessment.

TABLE 29.2

Clinical Behavioral Observations

Status and component	Observations
Physical	
Motor	Gait
	Upper extremity gross/fine motor movement
	Tone, spasticity, coordination, tremor
	Speech (articulatory control)
Sensory	Hearing
	Vision
	Perceptual disturbance (hallucinations)
Pain	Cautionary movement
	Muscle tension
	Grimacing
	Effect on emotional and cognitive status
	Effect on participation in therapies
Emotional	
Affect	Range, intensity, appropriateness to conversation
Effort	Level of effort in the assessment process including testing
Expression	Verbal and nonverbal
Regulation	Ability to monitor and regulate emotions
Cognitive	
Arousal level	Level of alertness, arousability
Attention	Span, divided, alternating
Memory	Encoding and retrieval processes
Language function	Word-finding problems, difficulty following commands
Thinking	Linear, circumstantial, tangential, perseverative
Ability to self-monitor	Impulsivity, disinhibition, insight
Visual–spatial	Unilateral neglect, visual field disruptions
Executive function	Novel problem solving, flexible thinking, impulse control

Rehabilitation Psychology Assessment Measures

Clinical evaluation in rehabilitation psychology includes a variety of structured, standardized assessment measures. Those measures used within the medical rehabilitation setting can be categorized using the Guiding Framework discussed earlier and outlined in Table 29.1. These component areas included in the Guiding Framework are discussed to offer a more comprehensive explanation of the importance of these areas in the assessment of rehabilitation populations. See Tables 29.3 through 29.5 for examples of measures from each of the Guiding Framework domains.

Psychological Functioning and Emotional Adjustment

Psychological functioning and emotional reaction to disability can have a profound effect on the medical rehabilitation process. Individuals participating in medical rehabilitation each arrive with different approaches to coping with distressing situations. Factors such as mood, anxiety, and personality characteristics can affect the rehabilitation process both positively and negatively. Identification of features that may have an adverse effect on rehabilitation can assist in tailoring interventions by the rehabilitation team to target the specific needs of the individual. For a listing of assessment measures relevant to aspects of psychological functioning and emotional adjustment, see Table 29.3.

TABLE 29.3

Psychological Assessment Measures in Rehabilitation Settings

Focus of assessment	Typical measures
Depression	Beck Depression Inventory—II (BDI–II)[a]
	BDI–Fast Screen[b]
	Center for Epidemiologic Studies Depression Scale (CESD)[c]
	CESD—Revised[d]
	Children's Depression Inventory[e]
	Geriatric Depression Rating Scale[f]
	Hospital Anxiety and Depression Scale (HADS)[g]
	Older Adult Health and Mood Questionnaire[h]
	Patient Health Questionnaire nine-item scale (PHQ-9)[i]
	Reynolds Adolescent Depression Scale[j]
	Reynolds Child Depression Scale[k]
	Stroke Inpatient Depression Inventory[l]
Anxiety	Beck Anxiety Inventory[m]
	Beck Youth Inventories for Children/Adolescents[n]
	Generalized Anxiety Disorder 7-item scale[o]
	Hamilton Anxiety Rating Scale[p]
	HADS[g]
	Reynolds Children's Manifest Anxiety Scale[q]
Personality	Millon Adolescent Clinical Inventory[r]
	Millon Clinical Multiaxial Inventory—III[s]
	MMPI—Adolescent version[t]
	Minnesota Multiphasic Personality Inventory—2[u]
	NEO Personality Inventory—3[v]
	Personality Assessment Inventory[w]
	The 16 Personality Factor Questionnaire[x]
Psychological distress	Brief Symptom Inventory[y]
	Millon Behavioral Medicine Diagnostic[z]
	PHQ[i]
	Short Form 36[aa]
	Symptom Checklist 90—Revised[bb]

[a]Beck et al. (1961). [b]Benedict, Fishman, McClellan, Bakshi, & Weinstock-Guttman (2003); Poole, Bramwell, & Murphy (2009). [c]Radloff (1977). [d]Eaton, Smith, Ybarra, Muntaner, & Tien (2004). [e]Kovacs (1982). [f]Jamison & Scogin (1992). [g]Zigmond & Snaith (1983). [h]Kemp & Adams (1995). [i]Kroenke, Spitzer, & Williams (2001). [j]Reynolds (2002). [k]Reynolds (1989). [l]Rybarczyk, Winemiller, Lazarus, Haut, & Hartman (1996). [m]Beck et al. (1988). [n]Beck, Beck, Jolly, & Steer (2005). [o]Spitzer et al. (2006). [p]Hamilton (1959). [q]Reynolds & Richmond (2008). [r]Millon (1993). [s]Millon, Davis, & Millon (1997). [t]Butcher et al. (1992). [u]Graham (2006). [v]McCrae & Costa (2010). [w]Morey (1991). [x]Cattell & Mead (2008). [y]Derogatis & Melisaratos (1983). [z]Millon et al. (2001). [aa]Ware et al. (1993). [bb]Derogatis et al. (1974).

Psychological distress. Global measures of psychological distress can be useful in a medical rehabilitation setting to capture information about a wide variety of psychological issues (for psychological assessment in general adult mental health settings and in medical settings, see Chapters 14 and 17 in this volume, respectively). The Patient Health Questionnaire (PHQ) is a self-report or provider-administered questionnaire that includes the nine-item mood module (PHQ-9) mentioned later as well as anxiety, alcohol, eating behaviors, and somatoform modules. The PHQ was developed from the Primary Care Evaluation of Mental Disorders and has been found to be a valid and reliable measure of the aforementioned areas of distress (Spitzer, Kroenke, & Williams, 1999).

The Short Form 36 contains eight health concepts derived from the Medical Outcomes Study

(Stewart & Ware, 1992) that represent several operational indicators of health including behavioral function, distress, well-being, and self-evaluations of general health status (Ware, Snow, Kosinski, & Gandek, 1993). The Millon Behavioral Medicine Diagnostic (MBMD) helps to identify qualities that affect treatment response and the likelihood of positive outcome from treatment (Millon, Antoni, Millon, Meagher, & Grossman, 2001). The MBMD includes three different normative samples including a sample of persons with chronic medical conditions, bariatric surgery candidates, and persons with chronic pain (Millon et al., 2001). It is a relatively easy measure to administer in rehabilitation settings, as it requires approximately 20 minutes to complete and can provide comprehensive information about an examinee's response to treatment that otherwise would require several measures (Millon et al., 2001).

The Symptom Checklist 90–Revised, (SCL-90-R; Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974), originally called the Hopkins Symptom Checklist, can be completed in less than 15 minutes and provides a measure of overall psychological distress, the intensity of symptoms, and the number of self-reported symptoms of psychopathology. A briefer measure that also provides a global psychological distress score and is based on the longer SCL-90-R is the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1983). As with the SCL-90-R, the BSI is a useful measure for assessing treatment progress. In addition to providing an overall global psychological distress score as well as information pertaining to depression and anxiety, it includes seven additional primary symptom dimensions including: somatization, obsessive–compulsive, interpersonal sensitivity, hostility, phobic anxiety, paranoid ideation, and psychoticism.

Depression. A variety of factors support the rationale for including measures of depression as part of routine assessment of individuals participating in medical rehabilitation. Depression is a serious condition that affects thoughts, emotions, physical health, behaviors, and relationships, and can have a significant effect on adjustment to medical conditions. Disturbance of mood can have an effect on a variety of areas including sleep, appetite, focus/concentration, pain perception, and ultimately one's ability to participate fully in medical rehabilitation. Depressive symptoms have been associated with decreased participation in rehabilitation therapies (Lenze et al., 2004) and increased propensity toward comorbid conditions (Baune, Adrian, Arolt, & Berger, 2006). Therefore, a thorough assessment of depressive symptoms is imperative in rehabilitation settings.

There is a disproportionately high prevalence of unipolar depression and dysthymic disorder among persons with chronic medical conditions (Baune et al., 2006), relative to the general population. This high proportion may be related to reaction to the situation and/or stress-related changes in neurochemistry. Depression has been associated with dysregulation of the hypothalamic–pituitary–adrenal (HPA) and sympathetic adrenal medullary axes (Appels, 1997). There is evidence to suggest that heightened activity of the sympathetic nervous system occurs in concert with dysregulation of the HPA axis, which may lead to insulin resistance and accumulation of visceral body fat (Björntorp & Rosmond, 2000), which is particularly detrimental among persons whose nervous systems already are under severe stress leading to metabolic complications (e.g., among persons with spinal cord injury [SCI]).

Depression includes a myriad of symptoms, which can be categorized either as cognitive–affective or somatic, depending on their presentation. Within medical settings, it can be difficult to ascertain whether somatic disturbances are related to depression or physical compromise associated with the underlying medical condition. Of note, cognitive–affective symptoms have been more closely associated with poor outcome (e.g., higher risk of cardiac-associated mortality) than somatic (Barefoot et al., 2000; Lésperance, Frasure-Smith, & Talajic, 1996).

Persons dealing with physical compromise and their providers may underestimate the link between their somatic complaints and depression, instead attributing somatic symptoms to recent surgical procedures and hospitalization (Lésperance & Frasure-Smith, 2000). However, at times, providers conclude that depression is present in individuals that are reporting symptoms that align with the somatic complaints of depression (e.g., sleep disturbance,

appetite changes, fatigue/reduced endurance, changes in sexual interest). More traditional measures of depression symptomatology, such as the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) tend to result in inflated scores for medical populations because several of the items assess somatic symptoms of depression. To more accurately assess the complex relationship between somatic symptoms of depression and somatic symptoms related to the medical condition, clinicians are encouraged not to rely solely on structured instruments of depression but rather to augment the assessment with a thorough review of symptoms during the clinical interview with the patient.

Given the concerns associated with distinguishing somatic symptoms of physical compromise from those associated with depression, rehabilitation psychologists have begun to explore measures that can more precisely assess acute symptoms of depression among persons participating in medical rehabilitation. One such measure, the Patient Health Questionnaire—9 (PHQ–9), is a nine-item depression scale that is derived from the more comprehensive PHQ. The items on the PHQ–9 are representative of the diagnostic criteria for depression in the *Diagnostic and Statistical Manual of Mental Disorders* (text revision; American Psychiatric Association [APA], 2000). The PHQ–9 is a short and easy-to-administer instrument designed to identify clinically manifest symptoms of a depressive episode and has been shown to be effective when used with medical populations (Gilbody, Richards, Brealey, & Hewitt, 2007; Kroenke, Spitzer, & Williams, 2001; Williams et al., 2005).

The Hospital Anxiety and Depression Scale (HADS) is also a popular choice in medical rehabilitation settings as it offers a concise method of operationalizing both depressive symptoms and anxiety in a medical environment (Zigmond & Snaith, 1983). Evidence of validity has been gathered for HADS scores and the HADS has been found to be a reliable measure of depressive and anxious symptoms, or overall psychological distress, across a variety of medical environments (Bjelland, Dahl, Haug, & Neckelmann, 2002; Spinhoven et al., 1997; Zigmond & Snaith, 1983).

Another option for measuring depression, particularly in older adults, is the Older Adult Health and Mood Questionnaire (OAHMQ; Kemp & Adams, 1995; for more information on psychological assessment with older adults, see Chapter 32, this volume). The OAHMQ is a 22-item measure that includes few somatic items, making it ideal for use with medically compromised individuals (Kemp & Adams, 1995). When compared with the PHQ-9, the OAHMQ tends to estimate a higher prevalence of depression (Krause et al., 2009). The OAHMQ has the ability to distinguish between nondepressed, clinically significant depressive symptoms, and probable major depressive disorder and has been found to be both a sensitive and specific measure of depressive symptoms among geriatric samples (Kemp & Adams, 1995) and medically compromised samples (Krause et al., 2009).

Anxiety. Anxiety can occur among individuals who are experiencing new-onset physical compromise and hospitalization, and it is very common in rehabilitation settings. Anxiety reactions can cooccur with sudden life transitions and after physical trauma (i.e., including traumatic injuries and acute or chronic medical conditions affecting motor, sensory, and cognitive function), and they often manifest when there is a perceived lack of personal control especially in novel situations, such as a rehabilitation setting. Somatic changes (e.g., increased heart rate, shortness of breath) can be misattributed by the patient and lead to anxious thoughts and emotions. This can have a severe effect on a variety of functional areas including physical, emotional, and interpersonal and can affect discharge planning.

Acute stress reactions and symptoms of acute stress (e.g., dissociative symptoms, re-experiencing, avoidance of stimuli that arouse recollections of the trauma, and hyperarousal) not only are common after a traumatic injury but also can manifest in relation to hospital-associated distress (e.g., intensive care unit; Jones et al., 2007; Schelling et al., 1998). Early assessment is imperative, and early intervention (i.e., treatment of targeted symptoms) is crucial for prevention of what may be considered "full-blown" posttraumatic stress disorder (Foa, Hearst-Ikeda, & Perry, 1995).

Assessment measures useful in identifying anxiety symptoms in rehabilitation settings include the HADS as described earlier (Zigmond & Snaith, 1983), the Beck Anxiety Inventory (Beck, Epstein, Brown, & Steer, 1988); the Hamilton Anxiety Rating Scale (Hamilton, 1959); and the Generalized Anxiety Disorder seven-item scale (Spitzer, Kroenke, Williams & Lowe, 2006). There exist few objective measures designed specifically for assessment of anxiety in medical rehabilitation populations. Further work in this area is indicated, especially with regard to brief measures designed to assess anxiety during early involvement in rehabilitation. This may assist rehabilitation psychologists and other treatment providers to target interventions that serve to enhance the rehabilitation process while also possibly reducing hospital length of stay (e.g., as with individuals whose anxiety may have an effect on their ability to wean from mechanical ventilation).

Personality. Personality characteristics can affect the rehabilitation process positively or can be detrimental and, as such, are important to assess in a medical rehabilitation setting (for more on assessment of personality and psychopathology, see Volume 1, Chapter 19, this handbook, and Chapters 10 and 11, this volume). Characteristics including resilience, extraversion, and cognitive flexibility have been shown to have a positive effect on coping during adversity (Campbell-Sills, Cohan, & Stein, 2006; Richardson, 2002) and, as such, may be related to positive outcome from rehabilitation. Other personality characteristics, such as neuroticism, may be detrimental (Campbell-Sills et al., 2006) and adversely affect the rehabilitation process, recovery, and social reintegration.

Several measures have been shown to be effective for characterizing personality factors in medical settings. A measure that may be useful for identifying aspects of personality associated with positive adjustment to disability is the NEO Personality Inventory—3 (NEO PI–3), which is an updated version of the Revised NEO Personality Inventory (NEO PI–R; McCrae & Costa, 2010). The NEO PI–3 is a 240-item measure of the five major domains of personality (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism; McCrae & Costa, 2010). The Personality Assessment Inventory (PAI; Morey, 1991) is a self-report measure consisting of 344 items that provide information pertaining to clinical (e.g., depression, anxiety), interpersonal (e.g., dominance), and treatment (e.g., suicidality) areas of consideration (Morey, 1991). The PAI is useful for identifying pathological aspects of personality as well as aspects associated with positive coping.

The Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Graham, 2006) is one of the most commonly used measures of personality in general clinical, research, and medical settings. It has been found to be a valid and reliable measure across a variety of patient groups as well as within general population samples, and the original MMPI was used often in medical settings (Piotrowski & Lubin, 1990). Support has been found for its utility in measuring normal personality patterns, patterns of behavior and coping, and screening for substance use disorders as well as for assessment of the psychological effects of medical conditions (Graham, 2006). In addition, the MMPI–2 can provide important information about an individual's propensity to engage fully in psychological and/or medical treatment (Graham, 2006).

Although the MMPI–2 is the most highly utilized instrument for assessment of psychopathology across a variety of samples, several drawbacks are associated with its use among persons participating in medical rehabilitation. Administration and scoring can be quite lengthy, as the MMPI–2 consists of 567 items with a variety of subscales requiring trained interpretation. In addition, cognitive disruptions, such as those associated with head trauma, can bias results and lead an examiner to erroneously attribute more emotional distress than that which is truly present because of difficulty differentiating emotional/personality symptoms from cognitive (Gass, 1991).

Health Behavior

Another important type of assessment that often is used in medical rehabilitation settings is the assessment of behaviors that serve to foster or hinder health maintenance and/or adherence to recommended health care management regimens.
A variety of behaviors can have an effect on the

rehabilitation process, including those related to pain, sleep, and necessary health behavior change. For a listing of relevant measures designed to assess key areas of health behavior, see Table 29.4.

Health behavior change. For many years, psychologists practicing in medical settings have been asked to assess and treat behaviors that exacerbate medical conditions and may lead to poor outcomes from medical rehabilitation. However, until recently it was difficult to bill for these types of assessments. In January of 2002, new billing codes for psychologists were developed through collaborative efforts of the APA's Practice Directorate and the Interdivisional Healthcare Committee to address the need for health and behavior (H&B) assessments. This is regarded as an important effort in recognizing psychologists as healthcare providers alongside their physician counterparts and has helped to establish psychologists as integral members of medical rehabilitation teams. Unlike psychodiagnostic assessments, H&B assessments primarily target behaviors identified as counterproductive to health maintenance and prevention of adverse medical consequences and do not focus on psychological or emotional factors (the reader also is referred to the chapter on psychological assessment in medical settings, Chapter 17 in this volume).

The Behavior Risk Factor Surveillance System (BRFSS) Questionnaire is a comprehensive measurement tool designed to assess a variety of health behaviors including physical activity, medical condition awareness, tobacco and substance use, and nutrition (Centers for Disease Control and Prevention, 2011). Although the BRFSS in its entirety is not something that would be administered in a medical rehabilitation setting, questions pertinent to health behaviors thought to have an effect on health status are available and can be used quite easily in rehabilitation settings.

Another type of assessment that often is indicated in medical rehabilitation settings is an assessment of substance use. Problematic use of a variety of substances including alcohol, illicit drugs, and opioid pain medications tends to be more prevalent among persons with disabilities than among the general population. Common screening measures designed to assess problematic alcohol consumption include the CAGE, which is an acronym for four items designed to provide a rapid assessment of alcohol abuse and dependence. This is a very brief and readily available measure that can be a part of any clinical assessment and can provide presumptive information pertaining to concerns for alcohol abuse or dependence (Ewing, 1984). The Alcohol Use Disorders Identification Test is a lengthier measure published by the World Health Organization that yields more comprehensive data pertaining to alcohol use behaviors (Babor, Higgins-Biddle, Saunders, & Monteiro, 2001).

Pain. Pain syndromes are common among individuals involved in medical rehabilitation and can have a serious effect on mood, sleep, and one's ability to participate in rehabilitation. Effective pain management with the use of medical (e.g., pain medications)

TABLE 29.4

Health Status and Management Assessment Measures

Focus of assessment	Typical measures
Fatigue	Multidimensional Fatigue Inventory[a]
	Fatigue Severity Scale[b]
Health behavior	Alcohol Use Disorders Identification Test[c]
	Behavioral Risk Factor Surveillance System (BRFSS)[d]
	Physical activity[e]
	Chronic health conditions[e]
Health behavior change	Alcohol consumption[e]
	Tobacco use[e]
	Exercise[e]
	Fruits and vegetables[e]
	CAGE Questionnaire of Alcohol Use (CAGE)[f]
Pain	Brief Pain Inventory (BPI)[g]
	McGill Pain Questionnaire (MPQ)[h]
	Minnesota Multiphasic Personality Inventory—2 (MMPI–2)[i]
	Numerical Rating Scale[j]
	Visual Analog Scale[j]
Sleep	Pittsburgh Sleep Quality Index[k]

[a]Smets, Garssen, Bonke, & De Haes (1995). [b]Krupp, LaRocca, Muir-Nash, & Steinberg (1989). [c]Babor et al. (2001). [d]Centers for Disease Control and Prevention (2011). [e]Subscale from the BRFSS. [f]Ewing (1984). [g]Cleeland & Ryan (1994). [h]Melzack (1975). [i]Epker & Block (2001). [j]Von Korff, Jensen, & Karoly (2000). [k]Fichtenberg et al. (2001).

and psychological strategies (e.g., cognitive–behavioral approaches) can optimize function and positively affect gains achieved in medical rehabilitation. The pain experience is related to a variety of factors including the underlying etiology as well as psychosocial factors including cognitive, emotional, and personality. At least a brief assessment of pain is imperative in most rehabilitation settings, including inpatient settings, and typically involves the patient's subjective report of location, intensity, type of pain, and indication of whether the pain interferes with performance of functional activities and/or participation in rehabilitation activities. This patient-reported information is coupled with behavioral observations made by the examiner and report of rehabilitation team members regarding the patient's participation in rehab activities. For persons with cognitive and/or communication impairments that preclude a more formal assessment of pain, observations of changes in facial expression, body posture, vocalizations, level and type of activity, and appetite can be informative with regard to pain experience and the effect of pain on function. Although a comprehensive assessment of pain and related syndromes is not typically necessary in an inpatient rehabilitation setting, it often is useful in outpatient settings (see Table 29.4) especially those designed to evaluate readiness for surgical procedures (e.g., spinal surgery or placement of spinal cord stimulators). Although it remains somewhat controversial, the MMPI–2 has been shown to be a useful instrument in assessing personality traits affecting the pain experience as well as outcome from spinal surgery (Epker & Block, 2001).

Sleep. It is important to assess quality and amount of sleep in medical rehabilitation settings because disturbances in sleep can have a serious effect on mood and daytime function and have been related to behavioral disturbances and difficulty with new learning (Fichtenberg, Putnam, Mann, Zafonte, & Millard, 2001). Sleep disturbances often arise in rehabilitation settings and can be related to changes in environment or schedules; emotional fluctuations and worries; psychological distress (e.g., nightmares associated with acute stress following trauma), and physical/medical factors. Sleep typically is assessed informally during the clinical interview. Although a formal assessment of sleep typically is not performed in inpatient medical rehabilitation settings, tools such as the Pittsburgh Sleep Quality Index (Fichtenberg et al., 2001) can be very useful in outpatient settings.

Fatigue. Fatigue is a well-documented and persistent consequence of many medical problems. Fatigue is a symptom experienced by a variety of rehabilitation populations, including those with traumatic brain injury (Ziino & Ponsford, 2006), multiple sclerosis (Bakshi, 2003), and cancer (Mitchell, Beck, Edwards Hood, Moore, & Tanner, 2007). Similar to the assessment of pain, a brief assessment of fatigue is important in most rehabilitation settings in order to determine if it is playing a role in the individual's ability to actively engage in the rehabilitation process and life activities. Evaluation involves the patient's subjective report of type of fatigue (physical and/or cognitive) and pattern of occurrence (e.g., diurnal variation, severity, impact on activity participation). Evaluation of sleep effectiveness, as well as psychiatric conditions, is important to consider in the presentation of fatigue because these have been found to cause, or at least contribute to, the fatigue experience (Labuz-Roszak, Kubicka-Bączyk, Pierzchała, Machowska-Majchrzak, & Skrzypek, 2012). However, fatigue may also occur in isolation from these other conditions and as a direct result of the injury, illness, or disease process.

Cognitive Functioning

Disruptions to cognition can have an effect on a variety of key areas of rehabilitation including one's awareness of cognitive impairments; ability to attend to information; one's ability to learn new information and skills; and one's ability to carry over and use skills learned in therapies across important domains. A neuropsychological evaluation allows the quantitative measurement of brain function, in terms of cognitive strengths and areas of difficulty, by an expert in brain–behavior relationships. Objective observations are made using standardized tests of cognitive ability across a variety of domains which are interpreted in light of the contextual variables associated with each patient's unique history. Table 29.5 includes examples of measures of neuropsychological functioning.

TABLE 29.5

Neuropsychological Assessment Measures in Adult and Pediatric Rehabilitation

Focus of assessment	Typical measures
Cognitive screenings	Mini-Mental State Examination[a]
	Montreal Cognitive Assessment[a]
	Neurobehavioral Cognitive Status Examination[a]
	Repeatable Battery for the Assessment of Neuropsychological Status[a]
Academic achievement and premorbid functioning	Bracken Basic Concept Scales—Revised National American[b]
	National American Adult Reading Test[a]
	Test of Nonverbal Intelligence[c]
	Wechsler Individualized Achievement Test[a]
	Wide Range Achievement Test—III[a]
	Woodcock–Johnson Psychoeducational Battery[c]
Infant development	Bayley Scales of Infant Development—III[b]
	Mullen Scales of Early Learning[b]
	Differential Ability Scales–Second edition[b]
	Stanford Binet Intelligence Scale–Fifth edition[b]
	Test of Nonverbal Intelligence–Second edition[c]
Intelligence	Wechsler Abbreviated Scales of Intelligence[c]
	Wechsler Adult Intelligence Scale—IV[a]
	Wechsler Intelligence Scale for Children—IV[b]
	Wechsler Preschool and Primary Scale of Intelligence—III[b]
Attention	Behavioral Inattention Test[a]
	Brief Test of Attention[c]
	Connor's Continuous Performance Test—II[c]
	Line Bisection Test[a]
	Paced Auditory Serial Addition Test[c]
	Test of Everyday Attention in Children[b]
	Trail Making Test[a]
Learning and memory	California Verbal Learning Test—II[c]
	Children's Memory Scale[b]
	Rey–Osterrieth Complex Figure Test[c]
	Wechsler Memory Scales[c]
	Wide Range Assessment of Memory and Learning[b]
Language	Boston Diagnostic Aphasia Examination[a]
	Boston Naming Test[c]
	Clinical Evaluation of Language Fundamentals[b]
	Controlled Oral Word Association Test[a]
	Multilingual Aphasia Examination—III[c]
Executive functions (planning, mental flexibility)	Word Fluency Test[a]
	Booklet Category Test[c]
	Delis–Kaplan Executive Function System[c]
	Halsted–Reitan Category Test[c]
	NEPSY Neuropsychological Battery–Second edition[b]
	Stroop Color Word Test (Adult or Child version) [c]
	Trails A and B (Adult or Child version)[c]
	Wisconsin Card Sorting Test[c]
Processing speed	Stroop Neuropsychological Screening Test[a]
	Symbol Digit Modalities Test[c]
Psychomotor	Finger Tapping Test[c]
	Grip Strength[c]
	Grooved Pegboard[c]
	Purdue Pegboard[b]

(*Continued*)

TABLE 29.5 (*Continued*)

Neuropsychological Assessment Measures in Adult and Pediatric Rehabilitation

Focus of assessment	Typical measures
Sensory, visual perceptual, visuospatial, and visuomotor	Clock Drawing[c]
	Hooper Visual Organization Test[c]
	Judgment of Line Orientation[c]
	Reitan–Klove Sensory Perceptual Examination[a]
Adaptive functioning	Adaptive Behavior Assessment Scale—Second Edition[c]
	Vineland Adaptive Behavior Scale—Second Edition[c]
Dementia scales	Consortium to Establish a Registry for Alzheimer's Disease[a]
	Mattis Dementia Rating Scale[a]

Note. All listed pediatric test measures were obtained from Baron (2004). All listed adult test measures were obtained from Lezak et al. (2004).

[a]Indicated only for adult testing. [b]Indicated only for pediatric testing. [c]Indicated for both pediatric and adult testing.

A mental status examination is performed with individuals who have suffered an injury or have a medical condition that potentially has an effect on cognitive functioning. The examination typically is brief and can include objective measures, such as the Mini-Mental State Examination (MMSE) which is designed to be a brief screen of the cognitive aspects of mental function (Folstein, Folstein, & McHugh, 1975). If the results of the MMSE raise concerns about the presence of acute cognitive decline, then typically more robust measures of neuropsychological functioning are administered to better characterize the pattern of impairment. It should be noted that the MMSE is a very gross screen of cognitive dysfunction and is not very sensitive to higher level cognitive decline. As such, follow-up in the outpatient setting for all individuals with brain-based injuries or illnesses, or who have complaints of cognitive decline, is standard care even when the individual performs well on the MMSE.

Neuropsychological assessment is important in understanding the nature and severity of central nervous system impairments presumed to underlie cognitive and behavioral disturbances (for more information on assessment in neuropsychological functioning, see Chapter 9, this volume). There has been a gradual shift in the role of neuropsychology from one of primarily differential diagnosis to greater emphasis on establishing the needs of the individual for neurorehabilitation interventions (Sohlberg & Mateer, 2001). The specific goals of neuropsychological services vary, but there are several goals common to all patients: (a) providing recommendations that will facilitate rehabilitation and adaptation to disability, including identifying areas of strength as well as barriers to acute and long-term recovery; (b) identifying potential functional impairments that can reasonably be inferred from neuropsychological data (e.g., inability to drive); (c) identifying potential safety issues associated with cognitive impairments (e.g., decision making); and (d) addressing factors that could adversely affect quality of life, including motivational and emotional issues. Comprehensive neuropsychological assessment has the potential to form the foundation for tailoring rehabilitation interventions. This occurs through characterization of the individual's cognitive–behaviorial strengths and difficulties; explaining the potential interplay between cognitive, emotional and physical factors in an individual's functioning; and directing the design of rehabilitation care to address the functional deficits rooted in the cognitive impairments.

There exist several hundred cognitive and neuropsychological tests, although not all have acceptable reliability and adequate evidence of validity (Lezak, Howieson, & Loring, 2004). The choice of assessment measures is based on several factors including the referral question; medical diagnosis; stability or instability of cognitive impairment; rehabilitation goals; and patient characteristics such as age or years of education.

In settings in which a brief determination of cognitive impairment must be made, such as an inpatient rehabilitation setting, brief cognitive screening measures typically are used, such as the Montreal Cognitive Assessment (www.mocatest.org), the Neurobehavioral Cognitive Status Examination (Lezak et al., 2004), or the Repeatable Battery for the Assessment of Neuropsychological Status (Lezak et al., 2004; see also Table 29.5). These generally sample multiple cognitive domains (e.g., orientation, attention, immediate and brief delay recall, and language), perhaps with one or a few measures, and are not intended to be used for diagnostic purposes. Screenings can be used to determine whether further evaluation (e.g., for diagnostic purposes) is warranted through a more comprehensive outpatient evaluation.

A comprehensive (i.e., outpatient) evaluation should focus on all major cognitive domains to make a definitive determination regarding cognitive integrity. This includes evaluating the (a) sensory and perceptual, (b) simple and complex attention, (c) processing speed, (d) learning and memory, (e) language, (f) motor, (g) visual–spatial organization, and (h) executive functions. Depending on the setting and population being served, neuropsychological evaluations also can include assessment of intellectual functioning, academic functioning, and vocational functioning. A comprehensive approach takes considerable time and attempts are made to ensure comprehensiveness and to adequately sample areas most likely to be impaired. Preferably, two or more instruments that measure the same domain are used to identify patterns of strength and impairment.

A variety of factors have the potential to have an effect on test performance and thus, an effect on results. Examinees must be able to sustain sufficient arousal to attempt completion of cognitive tasks. They must have adequate integrity of sensory functions to utilize and interact with test stimulus materials. For measures that require manipulation of test materials or writing, examinees must have adequate function of the upper limbs; otherwise, alternative measures must be chosen that do not require upper limb mobility and/or fine motor dexterity. Although nearly all psychometric tests of cognitive abilities have procedures for standardized administration, it can be appropriate to modify standardized test procedures for persons with disabilities while still preserving important aspects of the standardized administration (Heaton & Heaton, 1981; Hibbard, 1992). Psychologists might have an affirmative need to modify tests as part of a legal requirement for reasonable accommodations under the Americans with Disabilities Act (Ebener, Burkhead, & Merydith, 1994). Test selection may include measures that do not require manipulation of objects or writing for persons with upper limb motor problems (e.g., SCI); or those that include auditory stimuli for persons who are visually impaired or visually mediated stimuli for those who are hearing impaired (Roth et al., 1989). For a comprehensive list of tests that have had published modifications, see Caplan and Shechter (1995).

The examiner must have confidence that the examinee is putting forth her/his best effort because variable effort can render similar results to impaired performance. A variety of approaches have been developed to capture suboptimal effort, including both stand-alone instruments and embedded measures (see Boone, 2007, for a review). The pattern of performance is interpreted within the context of the evaluation including (a) examination of consistency in performance across tests that measure similar cognitive constructs; (b) examination of the discrepancy between test performance and severity of injury or illness; and (c) examination of the pattern of test performance and the "typical" pattern of deficits known to be common for the presumed underlying disorder. It is important to note that variable effort does not necessarily reflect intentional misrepresentation. The reasons for suboptimal effort vary and can range from mild detriments secondary to emotional states, such as depressed mood, to more pervasive detriments associated with conscious intentions to feign impairment.

Test interpretation is only as good as the normative information used to interpret the test performance. The most common demographic variables considered in compiling normative data are age and education, because these have been shown to be highly correlated with performance on cognitive tests (Lezak et al., 2004). In general, neuropsychological

tests that have a large normative sample (e.g., several hundred or more) and that include a sampling of diverse cultural, ethnic, gender, educational, intellectual, and age strata are the most likely to promote good inferences about test performance.

At times, members of the rehabilitation team may reach different conclusions regarding the patient's ability to function and accomplish required tasks. This can occur when the results of the neuropsychological evaluation and those from functional assessments differ (Chaytor & Schmitter-Edgecombe, 2003). It is important to note that declines in daily life functioning and difficulty in completion of functional tasks can be affected by several factors other than cognitive dysfunction including emotional distress, pain, fatigue, and reduced motivation. Performance for tasks that were previously learned, and had been performed often, tends to be more preserved compared with novel task performance after brain injury or illness. Neuropsychological and functional evaluations each contribute important, independent information about the patient's cognitive functioning and help to guide rehabilitation treatment planning with regard to rehabilitation goals, cognitive rehabilitation interventions, and need for supervision.

Specialty Areas for Consideration in Assessment

As noted throughout this chapter, rehabilitation psychologists often are asked to work with individuals with a variety of diagnoses and areas of impairment as well as to function in a variety of settings. They may be asked to provide an opinion pertaining to a patient's capacity to make an informed decision. In collaboration with vocational counselors, rehabilitation psychologists provide information pertaining to an individual's readiness and ability to return to work. Pediatric rehabilitation assessment focuses on the promotion of successful growth and development for children who are striving to regain functional skills. Each of these specialty areas for consideration in rehabilitation psychology assessment will now be described.

Decisional capacity assessment. There are several types of decision-making capacity, but the one most relevant and most frequently assessed in the medical rehabilitation setting is the ability to make reasonable medical decisions. The doctrine of informed consent requires clinicians to obtain voluntary and competent agreement to a medical intervention before performing the intervention and only after the patient has been informed of the material risks, benefits, and other facts of the condition and procedure. However, informed consent cannot be obtained from an individual who does not have the capacity to make an informed decision. The question of whether decision-making capacity is intact typically arises because of new onset cognitive impairments secondary to the medical condition, but it also may arise more subtly when a patient refuses to participate in rehabilitation interventions that are deemed by the treatment team to be in his or her best interest. It is important to note that, although the terms *decisional capacity* and *competency* often are used interchangeably, an important distinction is that decisional capacity is a clinical status (made by a clinician) of the ability to make decisions in a specific area based on a clinical assessment; whereas competency is a legal status (made by a judge) of the ability to retain decision-making power in a particular activity or set of activities (Baker, Lichtenberg, & Moye, 1998).

There are five main functional abilities inherent in demonstration of capacity: (a) *understanding* treatment situation and choices, (b) *evidencing* a treatment choice, (c) *appreciating* personal consequences of the choice, (d) *reasoning* about the treatment choice, and (e) *making* a reasonable treatment choice. Evaluation of medical decision-making capacity is a multistep process that includes clinical interview with patient and family members, garnering input from clinicians involved in the patient's care, and administration of capacity-specific measures. Several measures have been developed for use in decisional capacity assessment, and the MacArthur Competence Assessment Tools for Clinical Research and for Treatment is a standardized measure that appears to have the most empirical support for assessing an individual's decisional capacity (Dunn, Nowrangi, Palmer, Jeste, & Saks, 2006). The APA, in collaboration with the American Bar Association, published a guidebook on assessment of decision-making capacity, which is available online

free of charge (http://www.apa.org/pi/aging/programs/assessment/capacity-psychologist-handbook.pdf). Although the book is designed for the assessment of older individuals, the majority of the information included will assist any psychologist in a thorough assessment of medical decision making capacity in the rehabilitation setting.

Vocational assessment. Vocational assessment refers to the process of gathering relevant information about the individual's educational and occupational history followed by measuring relevant vocational variables. These variables typically include aptitudes, academic achievement, interests, personality, values, and needs. Once the measurement process has been completed, the psychologist or vocational counselor meets with the client, and perhaps family members, to integrate the results of the assessment into a vocational plan. When this assessment process occurs in the context of rehabilitation, the issue of the effect of the disability on the ability to return to the individual's former occupation or developing an alternative occupational path becomes the focus of the intervention. The goal is to outline accommodations and compensatory strategies that can assist the individual in being successful in the workplace (for more information on psychological assessment in work and organizational settings, see Volume 1, Chapter 22, this handbook).

Research has shown that unemployment is more prevalent among persons with disabilities as compared with their nondisabled peers despite developments in assistive technologies and innovations in rehabilitation approaches (Bruyère, Erickson, & VanLooy, 2004). As of 2009, the employment rate among community dwelling individuals aged 18–64 with a disability was 35% as compared with their nondisabled peers, 74% of whom were employed (Rehabilitation Research and Training Center on Disability Statistics and Demographics, 2010).

A vocational counselor often will work as a liaison between the client and a state vocational rehabilitation agency and will assist persons in inpatient and outpatient rehabilitation settings with vocational, situational, and return to work assessment (Bruyère et al., 2004). The vocational assessment ought to include a thorough evaluation of the individual's assets and vocational barriers including accommodations that will need to be made to account for physical, emotional, or cognitive challenges (Bruyère et al., 2004). A good understanding of the legal rights under the Americans with Disabilities Act (ADA) of individuals with disabilities and the legal responsibilities of their employers also can help to facilitate positive vocational placement and successful outcome (Bruyère et al., 2004). Vocational counselors and rehabilitation psychologists ought to have a thorough working knowledge of the ADA as well as any pertinent updates so they can provide the most relevant information and guidance to their clients. A comprehensive description of the ADA and requirements under the employment section of the Act can be found online (http://www.ada.gov).

Workplace accommodation needs will vary for each individual with a disability and will depend on the type and extent of physical/mobility, environmental, cognitive, and emotional effect. Therefore, it is essential that a thorough assessment of the individual's vocational needs is performed. This type of assessment typically includes an evaluation of the person's aptitudes and interests; an exploration of available job opportunities; identification of additional educational or training needs; and advocating for necessary accommodations (Rubin & Roessler, 2001). It also ought to include an assessment of intellectual, achievement, and neuropsychological functioning when working with an individual with a cognitive impairment who hopes to return to an educational or vocational setting.

When the results of the vocational assessment reveal that return to work is a reasonable goal and that further education or training is not needed, then the focus shifts to identification of a specific job of interest for which the person is qualified. The rehabilitation psychologist can assist with identifying vocational interests for career change. Standardized instruments useful in this regard include the Self-Directed Search (Donnay, Morris, Schaubhut, & Thompson, 2005) or the Strong Interest Inventory (SII; Donnay et al., 2005) and can assist with identifying interests that correspond with specific jobs. The SII was first published in 1927 and is the most thoroughly researched, respected, and used measure of vocational interests.

Pediatric assessment. Rehabilitation psychologists in a pediatric setting often are asked to assess patients with unique medical conditions and diagnostic groups, including developmental disorders such as spina bifida, cerebral palsy, genetic disorders, autism spectrum disorders, and attention-deficit/hyperactivity disorder. These conditions can present with abnormal developmental patterns as well as psychological and neuropsychological sequelae. Therefore, the effect of a new illness or injury may be different from that expected in adults or typically developing children. Extended hospitalization or illness can interfere with a child's development due to missed school and decreased opportunities for socialization.

Pediatric assessment requires a different set of assessment tools, depending on the age and developmental status of the child. Sattler (1992) noted that assessment of children frequently makes use of individually administered tests, behavioral observations, caregiver and teacher reports and checklists, and structured interviews. To evaluate a child, especially a young child, the environment must be adapted to be child friendly. Familiar toys, pictures, or family members should be present in the rehabilitation environment whenever possible. Because children often are unable to express symptoms, pain, or emotional distress verbally, behavioral or observational assessment tools are often helpful.

In children with medical conditions that have an effect on development, assessments that provide general developmental quotients or overall intellectual function (IQ) might not be the most useful in describing a child's abilities and impairments. Tests of intelligence, by virtue of being "global" measures, are prone to multiple confounding factors such as a child's language or motor deficits, which can have a differential effect on the overall score (Lezak et al., 2004). Because the measurement of intelligence does not become stable until approximately 4 years of age, tests of infant abilities are believed to be most useful for describing a child's current status, rather than having predictive value of a child's future cognitive ability, except for children with significant developmental delays (DuBose, 1977; McCall, Hogarty, & Hurlburt, 1972; Sattler, 1992). As such, careful and comprehensive assessment of the child's neuropsychological strengths and weaknesses, academic status, adaptive skills, motor skills and psychological status are likely to be more useful in educational and treatment planning for children with medical conditions (Barlow, 2001; Yeates, Ris, & Taylor, 1999).

School attendance is one of the primary tasks for children and adolescents. Careful evaluation of the psychological and cognitive factors that can have an effect on the success of the transition back to school is important. Rehabilitation psychologists frequently assist with recommendations for accommodations that facilitate successful integration or reintegration of a child into a school setting. Federal laws including The Education for All Handicapped Children Act (P.L. 94-142, now called the Individuals with Disabilities Act) provide mechanisms by which children can be provided with school services and accommodations (see also Volume 3, Chapter 1, this handbook).

Another concept unique to pediatric assessment is the idea that children may "grow into" deficits, long after an injury or illness has occurred (Holmes, 1987; Rudel, 1979). Younger children, for example, may not show deficits in executive function after a brain injury until much later in development, when the demands for these functions increase, and the brain may not have developed these abilities at the expected rate. As such, it is important to reassess children at critical developmental junctions (e.g., transitions to middle school and high school) to document changing patterns of strengths and weaknesses and make appropriate recommendations.

Mood disturbances often can occur after neurological injury, and are thought to arise from both the neurobiological changes related to the injury and the psychological adjustment to changes in abilities (Anderson, Catroppa, Haritou, Morse, & Rosenfeld, 2005; Felton & Revenson, 1984). These changes in mood and behavior can include internalizing behavior characterized by social withdrawal, anxiety, and depression and by externalizing behavior characterized by poor frustration tolerance, agitation, and verbal aggression. Children in need of rehabilitation have a greater risk for adjustment problems (Wallander & Thompson, 1995), and children with disabilities have been shown to have

more social and behavioral problems than children without disabilities (Cadman, Boyle, Szatmari & Offord, 1987; Werner & Smith, 1992; for more information on behavioral, social, and emotional assessment of children, see Volume 3, Chapter 6, this handbook).

The family environment and parental coping style have been shown to have an effect on functional status and social adjustment in a variety of developmental conditions and illness (Wallander et al., 1989a; Wallander, Varni, Babani, Banis, & Wilcox, 1989b). When working with children with traumatic brain injury, for example, knowledge of preinjury family cohesiveness and functioning may be helpful, as these factors are robust predictors of long-term outcome (Rivara et al., 1994; Wade, Taylor, Drotar, Stancin & Yeates 1996; Yeates et al., 1997). Psychological adjustment in the child often is closely tied with family factors such as cohesiveness, resilience, family support, maternal stress and coping, and parental reactions to injury or illness (Harper, 1984; Wells & Schwebel, 1987; Werner & Smith, 1992; for more information on assessment in family counseling, see Chapter 33, this volume).

SUMMARY

Rehabilitation psychologists have advanced training that prepares them for understanding the complex interactions among the make-up of the individual (psychiatric, emotional, behavioral, cognitive); the environment (rehabilitation setting, social/interpersonal, physical); and medical diagnoses, treatments, and related prognoses. Goals of assessment in medical rehabilitation settings include identifying components that may foster or hinder rehabilitation and recovery, all the while remaining mindful of the crucial role that the environment plays on individual function and quality of life. Assessment practices in the specialty field of rehabilitation psychology are based on a Guiding Framework designed to facilitate rehabilitation, recovery, and social reintegration. Through assessment, rehabilitation psychologists have the opportunity to play a vital role in enhancing function and quality of life for persons who are living with a disability.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Anderson, V. A., Catroppa, C., Haritou, F., Morse, S., & Rosenfeld, J. V. (2005). Identifying factors contributing to child and family outcome 30 months after traumatic brain injury in children. *Journal of Neurology, Neurosurgery, and Psychiatry*, *76,* 401–408. doi:10.1136/jnnp.2003.019174

Appels, A. (1997). Depression and coronary heart disease: Observations and questions. *Journal of Psychosomatic Research*, *43,* 443–452. doi:10.1016/S0022-3999(97)00158-X

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *The Alcohol Use Disorders Identification Test: Guidelines for use in primary care* (2nd ed.). Geneva, Switzerland: World Health Organization.

Baker, R. R., Lichtenberg, P. A., & Moye, J. (1998). A practice guideline for assessment of competency and capacity of the older adult. *Professional Psychology: Research and Practice*, *29*, 149–154. doi:10.1037/0735-7028.29.2.149

Bakshi, R. (2003). Fatigue associated with multiple sclerosis: Diagnosis, impact and management. *Multiple Sclerosis, 9*, 219–227.

Barefoot, J. C., Brummett, B. H., Helms, M. J., Mark, D. B., Siegler, I. C., & Williams, R. B. (2000). Depressive symptoms and survival of patients with coronary artery disease. *Psychosomatic Medicine, 62*, 790–795.

Barlow, D. H. (2001). *Clinical handbook of psychological disorders: A step-by-step treatment manual* (3rd ed.). New York, NY: Guilford Press.

Baron, I. S. (2004). *Neuropsychological evaluation of the child.* New York, NY: Oxford University Press.

Baune, B. T., Adrian, I., Arolt, V., & Berger, K. (2006). Associations between major depression, bipolar disorders, dysthymia, and cardiovascular diseases in the general adult population. *Psychotherapy and Psychosomatics, 75*, 319–326. doi:10.1159/000093955

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56,* 893–897. doi:10.1037/0022-006X.56.6.893

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571. doi:10.1001/archpsyc.1961.01710120031004

Beck, J. S., Beck, A. T., Jolly, J. B., & Steer, R. A. (2005). *Beck Youth Inventories for Children and Adolescents manual* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Benedict, R. H. B., Fishman, I., McClellan, M. M., Bakshi, R., & Weinstock-Guttman, B. (2003). Validity of the Beck Depression Inventory—Fast Screen in multiple sclerosis. *Multiple Sclerosis, 9*, 393–396. doi:10.1191/1352458503ms902oa

Bjelland, I., Dahl, A. A., Haug, T. T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression scale: An updated review. *Journal of Psychosomatic Research, 52*, 69–77. doi:10.1016/S0022-3999(01)00296-3

Björntorp, P., & Rosmond, R. (2000). Neuroendocrine abnormalities in visceral obesity. *International Journal of Obesity*, *24*(Suppl. 2), S80–S85. doi:10.1038/sj.ijo.0801285

Boone, K. B. (Ed.). (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. New York, NY: Guilford Press.

Bruyère, S. M., Erickson, W. A., & VanLooy, S. (2004). Comparative study of workplace policy and practices contributing to disability nondiscrimination. *Rehabilitation Psychology*, *49,* 28–38. HYPERLINK "http://dx.doi.org/10.1037/0090-5550.49.1.28" doi:10.1037/0090-5550.49.1.28

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Minnesota Multiphasic Personality Inventory—Adolescent: Manual for administration, scoring, and interpretation* (rev. ed.). Minneapolis: University of Minnesota Press.

Cadman, D., Boyle, M., Szatmari, P., & Offord, D. R. (1987). Chronic illness, disability and mental and social well-being: Findings of the Ontario Child Heath Study. *Pediatrics*, *79*, 805–813.

Campbell-Sills, L., Cohan, S. L., & Stein, M. B. (2006). Relationship of resilience to personality, coping, and psychiatric symptoms in young adults. *Behaviour Research and Therapy, 44*, 585–599. doi:10.1016/j.brat.2005.05.001

Caplan, B., & Shechter, J. (1995). The role of nonstandard neuropsychological assessment in rehabilitation: History, rationale, and examples. In L. A. Cushman & M. J. Scherer (Eds.), *Psychological assessment in medical rehabilitation* (pp. 359–391). Washington DC: American Psychological Association.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56*, 267–283. doi:10.1037/0022-3514.56.2.267

Cattell, H. E. P., & Mead, A. D. (2008). The Sixteen Personality Factor Questionnaire (16PF). In G. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The Sage handbook of personality theory and assessment: Vol. 2. Personality measurement and testing* (pp. 135–178). Los Angeles, CA: Sage.

Centers for Disease Control and Prevention. (2011). *Behavioral risk factor surveillance system survey questionnaire*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13,* 181–197. doi:10.1023/B:NERV.0000009483.91468.fb

Cleeland, C. S., & Ryan, K. M. (1994). Pain assessment: Global use of the Brief Pain Inventory. *Annals of the Academy of Medicine, Singapore*, *23,* 129–138.

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The Connor–Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*, 76–82. doi:10.1002/da.10113

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins symptom check-list (HSCL): A self-report symptom inventory. *Journal of Applied Behavioral Science*, *19,* 1–15.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13,* 595–605. doi:10.1017/S0033291700048017

Donnay, D. A., Morris, R., Schaubhut, N., & Thompson, R. (2005). Strong Interest Inventory manual: Research, development, and strategies for interpretation. Mountain View, CA: Consulting Psychology Press.

DuBose, R. F. (1977). Predictive value of infant intelligence scales with multiply handicapped children. *American Journal of Mental Deficiency, 81*, 388–390.

Dunn, D. S., & Dougherty, S. B. (2005). Prospects for a positive psychology of rehabilitation. *Rehabilitation Psychology*, *50,* 305–311. doi:10.1037/0090-5550.50.3.305

Dunn, L. B., Nowrangi, M. A., Palmer, B. W., Jeste, D. V., & Saks, E. R. (2006). Assessing decisional capacity for clinical research or treatment: A review of instruments. *American Journal of Psychiatry*, *163,* 1323–1334. doi:10.1176/appi.ajp.163.8.1323

Eaton, W. W., Smith, C., Ybarra, M., Muntaner, C., & Tien, A. (2004). Center for Epidemiologic Studies Depression Scale: Review and revision (CESD and CESD-R). In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Vol. 3. Instruments for adults* (3rd ed., pp. 363–377). Mahwah, NJ: Erlbaum.

Ebener, D. J., Burkhead, E. J., & Merydith, S. P. (1994). The Americans with Disabilities Act: Implications for vocational assessment. *Assessment in Rehabilitation and Exceptionality, 1*, 91–97.

Epker, J., & Block, A. R. (2001). Presurgical psychological screening in back pain patients: A review. *Clinical*

Journal of Pain, 17, 200–205. doi:10.1097/00002508-200109000-00003

Ewing, J. A. (1984). Detecting alcoholism: The CAGE Questionnaire. *JAMA, 252*, 1905–1907. doi:10.1001/jama.1984.03350140051025

Felton, B. J., & Revenson, T. A. (1984). Coping with chronic illness: A study of controllability and the influence of coping strategies on psychological adjustment. *Journal of Consulting and Clinical Psychology, 52*, 343–353. doi:10.1037/0022-006X.52.3.343

Fichtenberg, N. L., Putnam, S. H., Mann, N. R., Zafonte, R. D., & Millard, A. E. (2001). Insomnia screening in postacute traumatic brain injury: Utility and validity of the Pittsburgh Sleep Quality Index. *American Journal of Physical Medicine and Rehabilitation*, *80,* 339–345. doi:10.1097/00002060-200105000-00003

Foa, E. B., Hearst-Ikeda, D., & Perry, K. J. (1995). Evaluation of a brief cognitive-behavioral program for the prevention of chronic PTSD in recent assault victims. *Journal of Consulting and Clinical Psychology*, *63,* 948–955. doi:10.1037/0022-006X.63.6.948

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12,* 189–198. doi:10.1016/0022-3956(75)90026-6

Frank, R. G., & Caplan, B. (2010). Introduction. In R. G. Frank, M. Rosenthal, & B. Caplan (Eds.), *Handbook of rehabilitation psychology* (2nd ed.). Washington, DC: American Psychological Association.

Gass, C. S. (1991). MMPI–2 interpretation and closed-head injury: A correction factor. *Psychological Assessment*, *3,* 27–31.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, *22,* 1596–1602. doi:10.1007/s11606-007-0333-y

Graham, J. R. (2006). *MMPI–2: Assessing personality and psychopathology* (4th ed.). New York, NY: Oxford University Press.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32*, 50–55. doi:10.1111/j.2044-8341.1959.tb00467.x

Harper, D. C. (1984). Child behavior toward the parent: A factor analysis of mothers' reports of disabled children. *Journal of Autism and Developmental Disorders, 14*, 165–182. doi:10.1007/BF02409659

Heaton, S. R. H., & Heaton, R. K. (1981). Testing the impaired patient. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (526–544). New York, NY: Wiley.

Hibbard, M. R. (1992). The comprehensive psychological assessment of individuals with stroke. *NeuroRehabilitation, 2*, 9–20.

Holmes, J. M. (1987). Natural histories in learning disabilities: Neuropsychological difference/environmental demand. In S. J. Ceci (Ed.), *Handbook of cognitive, social, and neuropsychological aspects of learning disabilities* (pp. 305–319). Hillsdale, NJ: Erlbaum.

Jamison, C., & Scogin, F. (1992). Development of an interview-based geriatric depression rating scale. *International Journal of Aging and Human Development*, *35,* 193–204. doi:10.2190/0803-3FBC-6EB0-ACH4

Jones, C., Backman, C., Capuzzo, M., Flatten, H., Rylander, C., & Griffiths, R. D. (2007). Precipitants of post-traumatic stress disorder following intensive care: A hypothesis generating study of diversity in care. *Intensive Care Medicine, 33*, 978–985. doi:10.1007/s00134-007-0600-8

Kemp, B. J., & Adams, B. M. (1995). The older adult health and mood questionnaire: A measure of geriatric depressive disorder. *Journal of Geriatric Psychiatry and Neurology, 8*, 162–167.

Kortte, K. B., Gilbert, M., Gorman, P., & Wegener, S. T. (2010). Positive psychological variables in the prediction of life satisfaction following spinal cord injury. *Rehabilitation Psychology, 55*, 40–47. doi:10.1037/a0018624

Kovacs, M. (1982). *The Children's Depression Inventory: A self-rated depression scale of school-aged youngsters.* Pittsburgh, PA: University of Pittsburgh School of Medicine.

Krause, J. S., Saunders, L. L., Reed, K. S., Coker, J., Zhai, Y., & Johnson, E. (2009). Comparison of the Patient Health Questionnaire and the Older Adult Health and Mood Questionnaire for self-reported depressive symptoms after spinal cord injury. *Rehabilitation Psychology*, *54,* 440–448. doi:10.1037/a0017402

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*, 606–613. doi:10.1046/j.1525-1497.2001.016009606.x

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The Fatigue Severity Scale: Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology*, *46*, 1121–1123.

Labuz-Roszak, B., Kubicka-Bączyk, K., Pierzchała, K., Machowska-Majchrzak, A., & Skrzypek, M. (2012). Fatigue and its association with sleep disorders, depressive symptoms and anxiety in patients with multiple sclerosis. *Polish Journal of Neurology and Neurosurgery, 46*, 309–317.

Lenze, E. J., Munin, M. C., Dew, M. A., Rogers, J. C., Seligman, K., Mulsant, B. H., & Reynolds, C. F.

(2004). Adverse effects of depression and cognitive impairment on rehabilitation participation and recovery from hip fracture. *International Journal of Geriatric Psychiatry, 19*, 472–478. doi:10.1002/gps.1116

Lespérance, F., & Frasure-Smith, N. (2000). Depression in patients with cardiac disease: A practical review. *Journal of Psychosomatic Research, 48*, 379–391. doi:10.1016/S0022-3999(99)00102-6

Lésperance, F., Frasure-Smith, N., & Talajic, M. (1996). Major depression before and after myocardial infarction: Its nature and consequences. *Psychosomatic Medicine, 58*, 99–110.

Lewin, K. (1939). Field theory and experiment in social psychology: Concepts and methods. *American Journal of Sociology*, *44*, 868–896. doi:10.1086/218177

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment*. New York, NY: Oxford University Press.

Man, D. W. K., Lee, E. W. T., Tong, E. C. H., Yip, S. C. S., Lui, W. F., & Lam, C. S. (2004). Health service needs and quality of life assessment of individuals with brain injuries: A pilot cross-sectional study. *Brain Injury, 18*, 577–591. doi:10.1080/02699050310001646143

McCall, R. B., Hogarty, P. S., & Hurlburt, N. (1972). Transitions in infant sensorimotor development and the prediction of childhood IQ. *American Psychologist, 27*, 728–748. doi:10.1037/h0033148

McCrae, R. R., & Costa, P. T., Jr. (2010). *Professional manual for the NEO Inventories: NEO PI–3, NEO PI–R, and NEO FFI–3*. Odessa, FL: Psychological Assessment Resources.

Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain*, *1*, 277–299. doi:10.1016/0304-3959(75)90044-5

Millon, T. (1993). *Millon Adolescent Clinical Inventory: Manual*. Minneapolis, MN: National Computer Systems.

Millon, T., Antoni, M., Millon, C., Meagher, S., & Grossman, S. (2001). *Millon Behavioral Medicine Diagnostic*. Minneapolis, MN: NCS Assessments.

Millon, T., Davis, R., & Millon, C. (1997). *The MCMI–III Manual* (2nd ed.). Minneapolis, MN: National Computer Systems.

Mitchell, S. A., Beck, S. L., Edwards Hood, L., Moore, K., & Tanner, E. R. (2007), Putting evidence into practice: Evidence-based interventions for fatigue during and following cancer and its treatment. *Clinical Journal of Oncology Nursing, 11*, 99–113.

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Lutz, FL: Psychological Assessment Resources.

Piotrowski, C., & Lubin, B. (1990). Assessment practices of health psychologists: Survey of APA Division 38 practitioners. *Professional Psychology: Research and Practice, 21*, 99–106. doi:10.1037/0735-7028.21.2.99

Poole, H., Bramwell, R., & Murphy, P. (2009). The utility of the Beck Depression Inventory—Fast Screen in a pain clinic population. *European Journal of Pain*, *13*, 865–869. doi:10.1016/j.ejpain.2008.09.017

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1,* 385–401. doi:10.1177/014662167700100306

Rehabilitation Research and Training Center on Disability Statistics and Demographics. (2010). *Annual disability statistics compendium: 2010*. Retrieved from http://disabilitycompendium.org/pdf/Compendium2010.pdf

Reynolds, C. R., & Richmond, B. O. (2008). *The Revised Children's Manifest Anxiety Scale, second edition (RCMAS-2)*. Los Angeles, CA: Western Psychological Services.

Reynolds, W. M. (1989). *Reynolds Child Depression Scale (RCDS) professional manual*. Odessa, FL: Psychological Assessment Resources.

Reynolds, W. M. (2002). *Reynolds Adolescent Depression Scale—2nd edition (RADS–2): Professional manual*. Odessa, FL: Psychological Assessment Resources.

Richardson, G. E. (2002). The metatheory of resilience and resiliency. *Journal of Clinical Psychology, 58*, 307–321. doi:10.1002/jclp.10020

Rivara, J., Jaffe, K., Polissar, N., Fay, G., Martin, K., Shurtleff, H., & Liao, S. (1994). Family functioning and children's academic performance in the year following traumatic brain injury. *Archives of Physical Medicine and Rehabilitation, 75*, 369–379. doi:10.1016/0003-9993(94)90157-0

Roth, E., Davidoff, G., Thomas, P., Doljanac, R., Dijkers, M., Berent, S., . . . Yarkony, G. (1989). A controlled study of neuropsychological deficits in acute spinal cord injury patients. *Paraplegia*, *27*, 480–489. doi:10.1038/sc.1989.75

Rubin, S. E., & Roessler, R. T. (2001). *Foundations of the vocational rehabilitation process*. Austin, TX: PRO-ED.

Rudel, R. G. (1979). Neuroplasticity: Implications for development and education. In J. S. Chall & A. F. Mirsky (Eds.), *Education and the brain* (pp. 269–307). Chicago, IL: University of Chicago Press.

Rybarczyk, B., Winemiller, D. R., Lazarus, L. W., Haut, A., & Hartman, C. (1996). Validation of a depression screening measure for stroke inpatients. *American Journal of Geriatric Psychiatry, 4*, 131–139. doi:10.1097/00019442-199621420-00005

Sattler, J. M. (1992). *Assessment of children. Revised and updated* (3rd ed.). San Diego, CA: Author.

Schelling, G., Stoll, C., Haller, M., Briegel, J., Manert, W., Hummel, T., . . . Klaus, P. (1998). Health-related quality of life and posttraumatic stress disorder in survivors of the acute respiratory distress syndrome. *Critical Care Medicine*, *26,* 651–659. doi:10.1097/00003246-199804000-00011

Smets, E. M. A., Garssen, B., Bonke, B., & De Haes, J. C. J. M. (1995). The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research, 39*, 315–325.

Sohlberg, M. M., & Mateer, C. A. (2001). *Cognitive rehabilitation: An integrative neuropsychological approach*. New York, NY: Guilford Press.

Spinhoven, P, Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., & Van Hemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine, 27*, 363–370. doi:10.1017/S0033291796004382

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study. Primary evaluation of mental disorders. Patient Health Questionnaire. *JAMA*, *282,* 1737–1744. doi:10.1001/jama.282.18.1737

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine, 166*, 1092–1097. doi:10.1001/archinte.166.10.1092

Stewart, A. L., & Ware, J. E. (1992). *Measuring functioning and well-being: The Medical Outcomes Study approach*. Durham, NC: Duke University Press.

The Education for All Handicapped Children Act, P.L. 94-142 (1976).

Vellone, E., Rega, M. L., Galletti, C., & Cohen, M. Z. (2006). Hope and related variables in Italian cancer patients. *Cancer Nursing, 29*, 356–366. doi:10.1097/00002820-200609000-00002

Von Korff, M., Jensen, M. P., & Karoly, P. (2000). Assessing global pain severity by self-report in clinical and health services research. *Spine*, *25,* 3140–3151. doi:10.1097/00007632-200012150-00009

Wade, S. L., Taylor, H. G., Drotar, D., Stancin, T., & Yeates, K. (1996). Childhood traumatic brain injury: Initial impact on the family. *Journal of Learning Disabilities, 29*, 652–661. doi:10.1177/002221949602900609

Wallander, J. L., & Thompson, R. J. (1995). Psychosocial adjustment of children with chronic physical conditions. In M. C. Roberts (Ed.), *Handbook of pediatric psychology* (pp. 124–141). New York, NY: Guilford Press.

Wallander, J. L., Varni, J. W., Babani, L., Banis, H. T., DeHaan, C. B., & Wilcox, K. T. (1989a). Disability parameters, chronic strain, and adaptation of physically handicapped children and their mother. *Journal of Pediatric Psychiatry, 14*, 23–42. doi:10.1093/jpepsy/14.1.23

Wallander, J. L., Varni, J. W., Babani, L., Banis, H. T., & Wilcox, K. T. (1989b). Family resources as resistance factors for psychological maladjustment in chronically ill and handicapped children. *Journal of Pediatric Psychiatry, 14*, 157–173. doi:10.1093/jpepsy/14.2.157

Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey manual and interpretation guide*. Boston, MA: New England Medical Center, The Health Institute.

Wells, R. D., & Schwebel, A. (1987). Chronically ill children and their mothers: Predictors of resilience and vulnerability to hospitalization and surgical stress. *Journal of Developmental and Behavioral Pediatrics, 8*, 83–89.

Werner, E., & Smith, R. (1992). *Overcoming the odds: High-risk children from birth to adulthood.* Ithaca, NY: Cornell University Press.

Williams, L. S., Brizendine, E. J., Plue, L., Bakas, T., Tu, W., Hendrie, H., & Kroenke, K. (2005). Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke, 36*, 635–638. doi:10.1161/01.STR.0000155688.18207.33

Yeates, K. O., Ris, M. D., & Taylor, H. G. (1999). *Pediatric neuropsychology: Research, theory and practice*. New York, NY: Guilford Press.

Yeates, K. O., Taylor, H. G., Drotar, D., Wade, S. L., Klein, S., Stancin, T., & Schatschneider, C. (1997). Pre-injury family environment as a predictor of neurobehavioral outcomes following pediatric traumatic brain injury. *Journal of the International Neuropsychological Society*, *3,* 617–630.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*, 361–370. doi:10.1111/j.1600-0447.1983.tb09716.x

Ziino, C., & Ponsford, J. (2006). Selective attention deficits and subjective fatigue following traumatic brain injury. *Neuropsychology, 20*, 383–390.

CHAPTER 30

ASSESSMENT IN OCCUPATIONAL HEALTH PSYCHOLOGY

Jo-Ida C. Hansen

Occupational health psychology (OHP) is a relatively new specialty that emerged in the 1990s, through the joint effort of the American Psychological Association (APA) and the National Institute on Occupational Safety and Health (NIOSH). One of the early influences on the development of the OHP specialty was the Occupational Safety and Health Act of 1970 (P.L. 91-596), which helped to promote the recognition and control of job-related hazards. The Act authorized NIOSH to include motivational, behavioral, and psychological variables in research on worker safety and health and to evaluate the relation of job stress to illness and loss of work function (Cohen & Margolis, 1973). Momentum continued in 1974, when a special task force to the U.S. Secretary of Health, Education, and Welfare concluded that preventable workplace stressors result in sizable medical costs to employers and workers (Work in America, 1974). The dramatic increase in worker compensation claims for job stress, in the 1980s, then served as the tipping point for NIOSH to recognize stress-related psychological disorders as a leading occupational health risk (Sauter, Hurrell, Fox, Tetrick & Barling, 1999). The rate of lost work days ($Mdn = 25$) for workers with stress-related illnesses far exceeds that of work days lost because of other sources of nonfatal job illnesses or injuries ($Mdn =$ 6 days; Eggerth & Cunningham, 2012). The cost of stress to employers includes not only health care costs (nearly 50% greater for workers reporting high levels of stress) but also work-related injuries and turnover (Eggerth & Cunningham, 2012; Goetzel et al., 1998).

OHP, defined by NIOSH and APA as "the application of psychology to improving the quality of work life, and to protecting and promoting safety, health and well-being of workers" (NIOSH, 2004), is an interdisciplinary field that draws on many areas of psychology (e.g., vocational, counseling, industrial–organizational, social) as well as fields such as public health epidemiology, medicine, and human factors engineering. Advances in OHP are reflected in the wide array of research topics germane to the specialty including prevention, intervention, and treatment; work–family interactions; workplace aggression (e.g., violence, harassment, and bullying); organizational change (e.g., downsizing or restructuring); burnout; physical and emotional stress; job control and demands; workplace social support; cultural influences; technological advances, and globalization. In its broadest context, OHP is oriented toward (a) primary prevention using interventions targeted at the source of the problem, (b) secondary interventions that focus on workers' ability to respond to and cope with work-related problems, and (c) tertiary interventions that provide remediation (Houdmont & Leka, 2010).

OHP TESTING AND ASSESSMENT

Early assessment in the field often focused on physiological measures that captured variations in heart rate, respiration, perspiration, blood pressure, hormones, and muscle tension in reaction to stress. Psychological assessment of stress, on the other hand, focuses on affective responses such as anger, fear,

DOI: 10.1037/14048-030
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

depression, lethargy, distractibility, or inability to concentrate (Eggerth & Cunningham, 2012). OHP research has moved beyond its beginnings and its traditional focus on stress to incorporate many theoretical perspectives, qualitative and quantitative designs, and several approaches to assessment. Categories of frequently measured OHP variables include stressors; personality characteristics; situational characteristics in the environment; physiological, psychological, or behavioral strains that are the individual's response to stressors; and process and outcome measures used in OHP evaluations of interventions. Commonly used methods of assessment include interviews and focus groups, diary methods, and self-report instruments. Some lines of OHP research also use physiological measures. Interested readers may find other information on occupational attitudes in several of the chapters found in Volume 1 of this handbook on the topic of job satisfaction and other job-related attitudes (e.g., Chapters 28, 33, 35, and 37).

Physiological Measures

Work in occupational medicine, that served as a precursor to OHP, often focused on psychosomatic symptoms or physiological responses associated with workplace stressors. Evidence linked stressors with certain physical symptoms and diseases such as migraine headaches, sleep disturbances, cardiovascular disease, and diabetes (Kiecolt-Glaser, McGuire, Robles, & Glaser, 2002; Krantz & McCeney, 2002). The vast majority of current OHP research, however, focuses on psychosocial variables and infrequently employs physiological measures to assess outcomes, to substantiate self-reported affective stress, or to explain relations between stressors and disease. The most frequently used physiological measures include heart rate and blood pressure and gastrointestinal symptoms (Fried, Rowland, & Ferris, 1984).

Since the purpose of this chapter is to review psychological testing and assessment, detailed discussion of physiological measures will not be undertaken. A brief summary of biochemical assessment is offered simply to stimulate readers to consider the efficacy of employing psychophysiological biomarkers in their research. Endocrinological parameters provide an objective indication of job strain reactions and can complement the use of self-report strain measures. Studying endocrinological processes also can contribute to an understanding of the way in which job stressors may contribute to physical and psychological health problems (e.g., coronary disease).

Research has shown that stressors activate neural, neuroendocrine and endocrine pathways. Biological responses to stressful stimuli may be adaptive over short periods of time but continuous activation of biological defense mechanisms may cause health deterioration. Two systems—sympathetic-adrenal medullary (SAM) and hypothalamic–pituitary–adrenocortical (HPA)—are especially important in response to psychological stress. The SAM system influences epinephrine and norepinephrine secretion (catecholamines) and HPA influences secretion of glucocorticoids (Sonnentag & Fritz, 2006). Measurement of noradrenaline or peripheral adrenaline levels can assess activation of the sympathetic nervous system. Measures of heart rate or heart rate variability can assess the autonomic system function more generally. The measurement of cortisol fluctuations serves to provide a biomarker for HPA reactions to stress. Other physiological indicators of neuroendocrine reaction to stress include increases in prolactin and decreases in testosterone (Chandola, Heraclides, & Kumari, 2010; Chida & Steptoe, 2009; Hansen, Larsen, Rugulies, Garde, & Knudsen, 2009). Another area, for which physiological measures may be used, is assessing circulating immune cells such as NK (natural killer) cell markers, which play a role in common infection and in inflammatory diseases (Nakata, Takahashi, Irie, & Swanson, 2010), humoral immune markers such as immunoglobin G (IgG), and inflammatory cytokines such as interleukin-6 and tumor necrosis factor-α (Nakata, Takahashi, Otsuka, & Swanson, 2010).

Qualitative Methods of Data Collection for Assessment

Fortunately, relatively simple questionnaires that tap standard OHP variables (e.g., stressors, strain, well-being) provide sufficient information to test many hypotheses. However, on occasion research questions require more depth or nuance than

questionnaires can provide. In some instances, Likert-type statements may not allow participants to express completely why they are responding in one direction rather than another. In other instances, particularly in new areas of research, the questions asked by the research may not be sufficiently focused to allow the selection of a reasonable number of variables for study. In other cases (e.g., research within small organizations or focusing on underrepresented populations), the number of participants available may be insufficient to provide the sample sizes necessary for adequate power for the statistical analyses. Qualitative methods of assessment and data collection, including observation, focus groups, and interviews, often are used in these situations. Qualitative methods also are useful for identifying psychosocial stressors that are situation specific (Leka & Cox, 2010).

Observation. Observational methods of assessment may occur in field settings or in the laboratory. The technique involves either obtrusive observation (with the employee's knowledge) or unobtrusive observation (without knowledge) of the participant. In the case of OHP research, most often, the participants are employees and the observation occurs in their organizational setting. Observational data also can be obtained from videotapes from the environment; computerization allows analysis of behavior in real time and can include assessment of response duration and latencies (Tryon, 1998).

The effort needed to collect, code, and analyze observational assessments can be very time consuming and labor intensive. Also, the behavior of participants in obtrusive observation studies may be effected simply because the participants know they are under observation. On the occasion that unobtrusive observation occurs, ethical and legal requirements must be honored (Taris, deLange, & Kompier, 2010). An important ingredient in observational methods of assessment is careful attention to training the observers to record and code behaviors according to the study's protocol. Observer expectations or prejudice may result in bias; other sources of observer error include decay in observer performance over time, observer cheating (e.g., fabrication of data), and observer motivation (or lack thereof) to perform the task according to the protocol (Hartman, Barios, & Wood, 2004).

Focus groups and interviews. Focus groups, another method for qualitative data collection, use preselected participants who gather for a facilitated discussion. Interviews are similar to focus groups but typically involve one-on-one sessions with an interviewer and one or two interviewees (Bachiochi & Weiner, 2002). Focus group and interview assessments require careful identification of groups that should be present in the project. The number of sessions should match the point at which saturation is achieved (i.e., no new information is presented), but this endpoint can be difficult to predict. Typically, the group facilitator or interviewer has a script for guiding the session that includes an introduction, questions to be asked (usually in an open-ended format), and summary or closing information. A content analysis usually is done to provide a summary of the interviews or focus group discussions.

One drawback of focus groups and interview assessments, of course, is the extent to which the results cannot be generalized to a larger population. Conclusions based on the data by necessity must be conservative and confined to the specific setting for which the participants are representative. Similar to observational assessment, transcription of the discussions and interviews, as well as analysis of the content, is very labor intensive.

Self-Report Methods of Data Collection

Diary studies and self-report paper-and-pencil (or computer-administered) tests are the most frequently used approaches to psychological assessment and testing in OHP research.

Self-report diaries. Diary methods, which technically are self-report instruments, are structured to gather reports on experiences of participants' daily lives and to examine ongoing life experiences (Bolger, Davis, & Rafaeli, 2003; Wheeler & Reis, 1991). Diaries can be used to examine physiological, psychological, and social processes in real time (i.e., natural, spontaneous context) that minimize recall or retrospection errors (Tourangeau, Rips, & Rasinski, 2000). However, if within-person variability is expected to be small, retrospective accounts

most likely will be sufficiently accurate and real-time data collection is unnecessary.

Diary methods have utility for studying temporal dynamics. Whereas traditional longitudinal designs also can study change over time, diary methods—with more points of measurement—have greater fidelity for capturing changes. Diary methods also are used to study cyclical changes over time (e.g., circadian or diurnal rhythms or weekly, monthly or seasonal cycles).

Typically diary studies of within-person processes are nonexperimental. Using longitudinal data analysis methods to examine the temporal sequencing of events strengthens the inferences that can be made about cause and effect. An important step in designing studies that use diary methods is to determine the frequency and duration of assessments. In studies of high-frequency events, participants may be asked to report on only a certain number of instances. This approach requires appropriate participant training to reduce selection bias. Another choice for the researcher is to have participants report relevant events immediately or to allow them to postpone responses to an opportune time.

Fixed-time reporting schedules require careful consideration of the spacing intervals. Long intervals, for example, may mask natural cycles and may contribute to retrospection bias. Short intervals, on the other hand, may place too great a burden on participants and lead to noncompliance with the reporting schedule. Variable schedule or variable plus fixed schedule designs may help to reduce the potential for biased reports. However, this advantage needs to be weighed against the added burden on participants. Event-based designs (which are not mutually exclusive of time-based designs) require that the researcher provides a clear description of the triggering event to ensure that participants reliably identify each event.

Paper diaries are one of the most commonly used approaches to diary research. With this approach, participants are provided with a booklet or pack of questionnaires. However, honest forgetfulness can be a drawback, with participants failing to remember the scheduled response times or neglecting to have the diary with them. With technological advances, handheld and electronic devices now are an attractive option. The advantages are obvious. First, electronic data collection allows signaling the participant to report and also provides time or date stamps. Second, software can be programmed to provide flexibility in the presentation of questions, allowing randomization, and subsequent questions can be selected based on earlier responses. Third, electronic diaries simplify considerably the entry of data as well as reduce inaccuracies that result from the transcribing process. One caution in the use of electronic data collection methods is to avoid the assumption that all participants are facile with electronic technologies.

Diary methods are not without limitation. The use of diaries requires more prolonged training of participants than does the use of traditional survey designs and requires participant commitment for an extended time. Also, not much is known about the effect that diary completion may have on participants' responses. Concerns about both reactance and habituation are legitimate. Individual differences in personality motivation and ability also may have an effect on response compliance (Bolger et al., 2003).

Self-report questionnaires. The use of self-report questionnaires is, by far, the most popular method of testing and assessment in OHP research. One appeal, of course, is the relative ease with which information can be collected from a large sample of participants. With the development of online survey administration protocols, another advantage is the elimination of tedious data entry that may be prone to mistakes.

The quality of the instruments used in OHP research and practice ultimately determines how strong conclusions can be. Whereas some aspects of OHP assessment rely on well-established measures (e.g., personality assessment), the tendency in research settings has been to use ad hoc surveys or instruments developed for a single study, to modify existing measures (e.g., reduce the number of items or change the stimulus or response choice wording), or to combine portions of existing measures with minimal effort to establish evidence of validity or

provide reports of reliability that go beyond estimates of internal consistency. The trend toward refereed journals restricting the length of research reports also often means that short shrift is given to descriptions of the tests and inventories used in a study. This exacerbates the selection of sound instruments for future use (e.g., readers are not able to determine item content, item weights, or the extent to which scores have been normed or standardized) as well as interpretation of results (e.g., Is the outcome simply a function of measurement error?). Another potential source of measurement error results from the casual use of instruments developed on one national or cultural sample with a different national sample, with no effort to provide evidence of validity for cross-national or cross-cultural use of the test (see Volume 3, Chapter 26, this handbook).

Psychological assessment in OHP often is focused on stressors and strains. Job stressors are environmental variables that have a negative effect on people and have implications for psychological well-being and health. Strains are the reactions to stress and can include physical strains, behavioral strains, and psychological strains. As the field has evolved, the breadth of research topics under the OHP umbrella has expanded, and tests now commonly are used to measure symptoms of occupational stress (e.g., job dissatisfaction), the consequences of uncorrected job stress (e.g., burnout), personality characteristics (e.g., positive or negative affect), coping resources and responses (e.g., social support), work role characteristics (e.g., role ambiguity), job characteristics (e.g., shift work), relationships with work groups and supervisors (e.g., leadership style), organizational structure and culture (e.g., level within an organization), outcome evaluations (e.g., individual well-being or work climate), work–family interface (e.g., work–family conflict; WIF), and antisocial behaviors (e.g., bullying).

Psychological symptoms of occupational stress (i.e., strains) can include emotional reactions such as job dissatisfaction, depression, anxiety, boredom, and burnout. Behavioral symptoms of occupational stress (i.e., strains) also are frequently assessed and include employee behaviors such as alcohol and drug use, absenteeism, and accident proneness.

Factors that exert an influence on workers' levels of well-being (i.e. strains) include work demands, working hours, and job control (O'Driscoll & Brough, 2010). Work demands include both the amount of work required and the time allotted for task completion. Role ambiguity is uncertainty about the way in which to perform a job, and role conflict occurs when a person faces incompatible job demands. Role overload represents an excessive number of roles that a person must fulfill. Job control is the extent to which people can have autonomy or personal control over their work vis-á-vis decision authority, prioritization, and how to do the job or task (Cooper, Dewe, & O'Driscoll, 2001).

In some instances, factors may serve either as a stressor or as a buffer between stressors and strains. One example is social support, which, broadly defined in the OHP literature, includes social support at work, from supervisors and colleagues; and outside work, from family and friends. The support may include mentoring, informational support, practical help, and feedback. Lack of social support, however, can be a stressor that leads to reduced psychological health. On the other hand, research has shown that the presence of social support protects people from unhealthy work environments.

Other factors, such as work–family interface, may be either a stressor or a strain. High levels of family–work conflict (FIW) may have a negative effect on people's work attitudes and performance. On the other hand, work demand may contribute to strain that emerges in the form of decreased marital and family satisfaction for the worker, spouses and partners, and children. Antisocial behaviors also may be stressors or may be symptoms of strain. Workplace harassment, bullying, and violence and aggression can lead to anxiety, fear, depression, and reduced performance and turnover. However, workers experiencing workload stressors or those who are in low-control positions may exhibit antisocial behaviors.

The number of instruments available for OHP research and practice is huge, and the instruments used often overlap with those selected in other

domains (e.g., measures of personality that cut across counseling, industrial–organizational, clinical, sports, and occupational health psychology). This review focuses to some extent on exemplars of measures and attempts to highlight those instruments most likely not reviewed in other chapters. In many instances, the instruments were designed solely for research purposes, and their authors have not assigned formal names to the scales or tests as one might do with a commercial product.

Job stress. Traditionally, the measures included in the category of job stress or stressors assessment often focus on job characteristics or job roles and organizational variables. More recently, this category also has included variables external to the work environment that might cause feelings of strain related to work.

Potential workplace stressors can be divided into two clusters: workplace characteristics that relate to the role of the worker and those that relate to the organization and to the organization's culture. Within this context, the employee's perception of the job has become increasingly important in OHP research, and self-report questionnaires provide an efficient method for gathering incumbent workers' evaluations of their jobs and their organizations.

Job roles. The most frequently studied job roles include role ambiguity, role conflict, role overload, and autonomy. *Role ambiguity* is an uncertainty about the way in which to perform a job. Breaugh and Colihan (1994) developed a nine-item measure of role ambiguity to assess three facets: (a) ambiguity about methods to perform the job, (b) uncertainty about work scheduling (e.g., sequencing and time allocation), and (c) ambiguity about standards for measuring employee performance. Test–retest reliabilities are most robust for the performance scale (.80) and smaller for scheduling (.73) and work method ambiguity (.65). Coefficient alpha values range from .80 to .97.

Role conflict can be the incompatibility of demands of a job or incompatibility between the expectations of the employee and employer (or worker and supervisor). Rizzo, House, and Lirtzman (1970) developed one of the first measures to capture role conflict as well as role ambiguity. Several studies (Fields, 2002) have explored the internal consistency of the two facets (range = .71–.87 for role conflict and .71–.95 for role ambiguity), and evidence of validity supports the conclusion that the two facets are distinct constructs (Netemeyer, Johnston, & Burton, 1990). In one study, the correlation between role conflict and role ambiguity was found to be .25 (Rizzo et al., 1970).

In contrast to the Rizzo et al. (1970) measure, which assesses only negative aspects of conflict and ambiguity, House, Schuler, and Levanoni (1983) contributed a measure that attempted to also capture the positive aspects of role conflict and role ambiguity. Their effort was partially successful. The role ambiguity factor included seven positive items and four negative items. However, the factor that emerged for role conflict contained seven items that measured only stress caused by other parties. Coefficient alphas range from .79 to .86, and the two scales correlate positively with psychological strain and distress and negatively with job satisfaction (Fields, 2002).

Role overload is when a person must fulfill too many roles or do too much work in the time available. Bacharach, Bamberger, and Conley (1990) and Beehr, Walsh, and Taber (1976) each developed three-item scales to assess work overload. The Bacharach et al. measure focuses on the facet related to having time to do the job. Coefficient alpha values range from .60 to .64, and the scale scores correlate positively with role conflict and family conflict and negatively with team efficacy (Fields, 2002). The Beehr et al. scale taps both performance standards and time to do the job. Scale scores correlate positively with job dissatisfaction, tension, and fatigue, and scores on the scale have an internal consistency value of .71 (Beehr et al., 1976).

Another variant of role overload is workload that is measured with a variety of approaches, including the amount of work, the pace of the work, and the mental demands (or lack thereof) of the work. A Dutch-language instrument developed by Van Veldhoven and Meijman (1994) and translated into English by Van Yperen and Snijders (2000), is often adapted for use in OHP research. The 11-item instrument assesses both the amount of work as well as the pace. Alpha coefficients range from .71 (Van Veldhoven, Broeresen, & Fortuin, 1999) to .86 (Van Yperen & Snijders, 2000).

Caplan, Cobb, French, Van Harrison, and Pinneau (1980) also developed the Job Overload Inventory (11 items) to assess an employee's perceived pace and amount of work. Coefficient alpha values range from .72 to .81, and the scale scores correlate positively with hours worked and negatively with work satisfaction (Fields, 2002).

Spector and Jex (1998) developed the Quantitative Workload Inventory (QWI), which assesses how often a job requires various types of work. They report a mean coefficient alpha of .82 across 15 studies using the measure. The QWI scores correlate .38 with role conflict and .33 with hours worked per week.

Control over, and autonomy in, one's work environment is another work role construct of importance to OHP models of job stress. *Control* is when a worker can influence the environment (Ganster, 1989), and *autonomy* is the extent to which employees have discretion over their work schedules, sequencing, and approaches to job tasks. In the context of his Demands-Control model of job stress, Karasek (1979) developed the Job Demand (seven items) and Decision Latitude (eight items) Scales. The Work Demand Scale scores correlate positively with distress, hours worked per week, and somatic complaints and negatively with social support from managers and job satisfaction. Coefficient alpha values range from .79 to .88. The Decision Latitude Scale scores correlate positively with job level, hours worked per week, and job satisfaction and negatively with somatic complaints. Alpha coefficient values range from .77 to .85 (Fields, 2002). Dwyer and Ganster (1991) developed a 22-item Work Control Scale that assesses four facets: control over (a) procedures and policies, (b) order of task performance, (c) break schedules, and (d) physical work environment. Scores on the Work Control Scale correlate positively with decision authority and work satisfaction, and the internal consistency value is .87 (Connell, Lee, & Spector, 2004).

Spector and Fox (2003) developed the Factual Autonomy Scale to replace the subjectivity that is prevalent in other measures of autonomy with fact-based questions. Correlations between worker self-report and supervisor ratings of the work situation, and between self report and coworker ratings, are .53 and .38, respectively, and the alpha coefficient value is .81. Factual Autonomy Scale scores correlate .42 with the Job Diagnostic Survey autonomy scale (Hackman & Oldman, 1975).

Monotony of the work environment is another employee perception that is important in evaluations of work. Melamed, Ben-Avi, Luz, and Green (1995) developed the four-item Subjective Monotony Scale to measure employee perceptions of the monotony of their work. The scale scores correlate negatively with education, age, and job satisfaction and positively with psychological distress, anxiety, and somatic complaints. The coefficient alpha values range from .68 to .76 (Fields, 2002).

Organizational variables. Job stressors that relate to the organization rather than to the employee's role include organizational variables such as justice or fairness, workload, career development variables, and physical work environment. The Physical Work Environment Satisfaction Questionnaire (Carlopio, 1996) assesses five facets: design of the physical environment (e.g., air quality and lighting), work and system characteristics (e.g., information availability and work pace), plant facilities (e.g., eating facilities, cleanliness, office size), health and safety (e.g., hazards control, training), and equipment design (e.g., tools, machines, and materials). Alpha coefficients for the subscales range from .71 to .93. Scores for the overall scale correlate with organizational commitment (.54), intentions to leave the organization (−.44), and work satisfaction (.65; Carlopio, 1996).

The Career Aspiration Scale (CAS; O'Brien, 1996) measures three facets typical of people who aspire to advance in their careers: (a) leadership and promotion, (b) training and managing others, and (c) pursuing additional education. Evidence of reliability and validity for the CAS has been established only for women and girls. A series of factor analyses suggests two CAS factors: (a) Leadership and (b) Achievement Aspirations and Educational Aspirations. Two week test–retest stability coefficients are .84 and .71 for Leadership and Achievement and Educational Aspirations factors, respectively. Alpha coefficient values range from .63 to .82 (Gray & O'Brian, 2007).

Several measures have been developed to assess two components of organizational justice or fairness: procedural justice, which focuses on the way in which decisions are made; and distributive justice, which focuses on a summary judgment about the fairness of decisions (Fields, 2002). Sweeney and McFarlin (1997) constructed a 13-item procedural justice subscale that assesses fairness of organizational procedures including promotion, performance feedback, and solving work-related problems. They also developed an 11-item distributive justice subscale that assesses perceived fairness in awarding rewards such as raises, general recognition, and promotions. The coefficient alpha for the Procedural Justice scale is .84 and .81 for the Distributive Justice scale. A confirmatory factor analysis supported the hypothesis that the two scales are empirically distinct. Scale scores for both subscales correlated positively with job satisfaction, pay level, organizational commitment, and intention to stay in a job (Sweeney & McFarlin, 1997).

Niehoff and Moorman (1993) developed subscales to measure distributive and procedural justice and also to measure interactive justice, or the extent to which employees perceive that their needs are considered in decision making. Confirmatory factor analysis found that the three scales are empirically distinct. Interactive justice scores correlate positively with formal meetings, and scores on all three subscales correlate positively with altruism, courtesy, sportsmanship, conscientiousness, and civic virtue (Niehoff & Moorman, 1993). The coefficient alphas are .72 for distributive justice, .85 for procedural justice, and .92 for interactive justice.

A Perceived Injustice Scale was developed by Hodson, Creighton, Jamison, Rieble, and Welsh (1994). This measure assesses the degree to which the employee agrees with four specific statements about employers treating employees unfairly. The coefficient alpha value is .70, and the scale scores correlate positively with working in physically demanding jobs and large organizations and negatively with higher socioeconomic status and bureaucratic procedures.

Broad assessments of job stressors. The Job Stress Survey (Spielberger, 1994) is a comprehensive 30-item instrument designed to measure perceived intensity (severity) and frequency of occurrence of working conditions that have a negative effect on the psychological well-being of workers. Internal consistency values range from .87 to .93 (Spielberger & Reheiser, 1995). The Job Stress Survey includes two subscales: Job Pressure and Organizational Support.

Another broad measure of job stress is the Generic Job Stress Questionnaire (GJSQ; Hurrell & McLaney, 1988) developed by the National Institute of Occupational Safety and Health. The GJSQ has 13 scales that measure job stressors (e.g., role demands, skill underutilization, job control, conflict, workload) and a number of scales to measure strain (e.g., depression, illnesses, job dissatisfaction, somatic complaints) and modifier–mediator variables (social support and self-esteem). Internal consistency reliabilities range from .75 to .89. The GSJQ has been translated into several languages (e.g., Finnish and Japanese; Hurrell, Nelsen, & Simmons, 1998).

Strain. Strain is the psychological, behavioral, or physical reaction that people have to job stress or job stressors. Many strain measures have been developed to assess reactions to stressors in life, and often these measures are adapted for use in the workplace. Emotional reactions to stress, for example, may include anger, anxiety, depression, psychopathology, and well-being (or lack thereof). Scales also exist to assess physical reactions to stress such as headaches, stomach distress, and allergic reactions. Work attitudes such as job satisfaction, job affect, and organizational commitment, and workplace behaviors such as absences, turnover, and antisocial behaviors are constructs frequently assessed as workplace-specific markers of strain.

Work attitudes. Examples of satisfaction measures include the Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969), which measures five facets of job satisfaction—work, pay, supervision, coworkers, and promotion opportunities; the Minnesota Satisfaction Questionnaire (MSQ; Weiss, Dawis, Lofquist, & England, 1966), which measures 20 facets; and the Job in General Scale (Ironson, Smith, Brannick, Gibson, & Paul, 1989). The 72-item JDI (Smith et al., 1969) was updated in 1989 by M. Roznowski. Coefficient alpha values for the five

scales range from .75 to .94 (Fields, 2002). Factor analysis of the JDI (Roznowski, 1989) confirmed that the items load on five distinct factors. The MSQ (Weiss et al., 1966) is available in a 20-item form that measures general job satisfaction (coefficient alpha values range from .85 to .91; Fields, 2002) and in a 100-item form that includes 20 satisfaction scales. Factor analyses of the MSQ have found both two-factor (intrinsic and extrinsic satisfaction) and four-factor solutions (intrinsic and extrinsic satisfaction, recognition and authority/sociability; Fields, 2002). The Job in General Scale (Ironson et al., 1989) is designed to assess global job satisfaction. The coefficient alpha values for the Job in General Scale range from .82 to .94 (Fields, 2002). Scale scores correlate positively with affective organizational commitment, trust in management, judgments of fairness, and tenure with a supervisor.

Greenhaus, Parasuraman, and Wormly (1990) developed a career satisfaction scale that measures satisfaction with career success. The coefficient alpha values for the five-item scale range from .83 to .89. The scale scores correlate positively with salary, promotions received, perceptions of upward mobility, supervising support, and person–organization value congruence (Fields, 2002).

The Job-Related Affective Well-Being Scale (JAWS; Van Katwyk, Fox, Spector, & Kelloway, 2000) was constructed to measure a wide range of emotional reactions to work. The instrument captures two dimensions: a pleasure–displeasure dimension and a low–high arousal dimension. The instrument includes five scales measuring overall well-being, high arousal, low arousal, pleasure, and displeasure score. The alpha coefficients range from .80 (low pleasure and low arousal) to .95 (JAWS composite). The JAWS scale scores correlate as expected with measures of physical symptoms, intentions to quit a job, work interpersonal conflict, and workload.

Measures of organizational commitment assess the strength of an individual's identification with an organization. The Organizational Commitment Questionnaire (OCQ; Mowday, Steers, & Porter, 1979) is a measure of overall commitment. The internal consistency reliability for the OCQ averages .80 over 90 samples (Mathieu & Zajac, 1990), and the scale has been shown to be distinct from work ethic and job involvement (Marrow, 1993). The 15-item OCQ primarily measures affective commitment, and its items load on the same factor as do the items on Allen and Meyer's Affective Continuance Scale (1990; Shore & Tetrick, 1991), one of three scales they developed to measure commitment; the other two scales are Continuance and Normative Commitment.

Meyer and Allen's (1997) short version of their instrument has six items to measure each type of commitment. Coefficient alpha values range from .77 to .88 for Affective Commitment (AC), from .65 to .86 for Normative Commitment (NC) and from .69 to .84 for Continuance Commitment (CC). Some studies suggest that the CC Scale contains two related dimensions – one reflecting personal sacrifice and the other lack of alternatives. Also, the level of the relations between AC Scale scores and NC Scale scores suggests that affective commitment and obligation to an organization are not independent of one another (Meyer, 1997).

Burnout. One of the most well-known strain constructs is *burnout*: "A state of physical, emotional and mental exhaustion caused by long term involvement in emotionally demanding situations" (Pines & Aronson, 1988, p. 9). Maslach developed the Maslach Burnout Inventory (MBI; Maslach & Jackson, 1986) to measure the frequency and the intensity of the three dimensions of burnout: emotional exhaustion, depersonalization, and reduced personal accomplishment. Internal consistency coefficients for the six subscales range from .71 to .90 (Hurrell, Nelson, & Simmons, 1998). The evidence of validity for scales representing the three factors has been studied repeatedly, and most results support the three-factor solution (Shirom, 2010). However, the emotional exhaustion dimension is a core component of the MBI and has been found to predict both cynicism and reduced personal effectiveness. The MBI–General Survey (MBI-GS; Maslach, Jackson, & Leiter, 1996) was developed to provide an instrument less tied to human service occupations than was the original MBI. The MBI-GS measures a more general type of exhaustion (Exhaustion Scale) than does the MBI, and the Cynicism

Scale replaces the Depersonalization Scale of the MBI. The Professional Efficacy Scale is similar to the Personal Accomplishment Scale of the MBI. Exhaustion has been shown to predict absenteeism, and cynicism and efficacy have been shown to predict turnover and satisfaction (Bakker, Demerouti, & Schaufeli, 2002).

Some argue that professional efficacy, as proposed and measured by Maslach, is an outcome, rather than a facet, of burnout. The German-language Oldenburg Burnout Inventory (OLBI; Dermerouti, Bakker, Vardakou, & Kantas, 2003) was developed to focus only on the core dimensions of burnout: exhaustion and disengagement (from work). The OLBI Exhaustion scale includes items related to affective, physical and cognitive aspects of exhaustion, which expands the use of the OLBI to include workers who perform physical work in addition to those whose jobs are mainly to process information. The OLBI items assess distancing oneself from one's work in general and from work content (Demerouti & Bakker, 2008). Initial evidence of validity for the scale scores of the OLBI used factor analysis with samples drawn from Germany, Greece, and the United States. In all cases, the results showed a two-factor structure: exhaustion and disengagement. Scores for parallel scales of the OLBI and MBI correlate .48 or higher, and test–retest coefficients over 4 months are .51 for Exhaustion and .34 for Disengagement. Coefficient alpha levels for both scales are .85 (Demerouti & Bakker, 2008). The OLBI items include positively and negatively worded items (i.e., items assessing exhaustion and vigor as well as disengagement and dedication), which allows the OLBI to be used to assess burnout and work engagement simultaneously. Scales for an English translation of the OLBI, created by Evangelica Demerouti using back-translation methods, have internal consistency coefficients that range from .74 to .87. Test–retest score reliabilities for the English version range from .34 to .51. The two-factor model also is the best fit for the English version (Halbesleben & Demerouti, 2005).

Work behaviors. The strain that results from stress can be visible in work behaviors, some of which can have consequences for the worker (e.g., aggression toward others) and some of which can have consequences for the organization (absenteeism, accident proneness, leaving the job). Several instruments are available to assess the effect of illness on productivity. One of the most extensively developed work behavior assessments is the Work Productivity and Activity Impairment Questionnaire (WPAI; Reilly, Zbrozek, & Dukes, 1993). One advantage of the WPAI is that the instrument is generic and does not contain questions specific to a type of illness or type of employment. The five items assess hours absent from work because of (a) health problems and (b) other reasons as well as (c) hours actually worked, (d) productivity at work, and (e) daily activity other than work. Test–retest correlation coefficients range from .71 for overall productivity at work to .87 for productivity outside of work. Scores on the WPAI correlate with general health perceptions, pain, and global measures of work and interference with regular activities (Prasad, Wahlquist, Shikiar, & Shih, 2004). The Health and Work Performance Questionnaire (HPQ; which Kessler et al. [2003] developed with the World Health Organization), measures four work functions: absenteeism, performance at work, work-related accidents or injuries, and job turnover. Good agreement has been found between scores on the HPQ and company archival data.

Some instruments have been developed to assess the effect of specific illnesses on productivity. The Migraine Disability Assessment Questionnaire (MIDAS; Stewart, Lipton, Kolodner, Liberman, & Sawyer, 1999) is one example. The MIDAS measures the effect of migraines on paid labor, household chores, and other daily activities (e.g., leisure activities). The test–retest coefficient is .75, and Cronbach's alpha is 0.83.

Other instruments have been developed to assess on-the-job behaviors that can range from antagonistic to positive. The 12-item Employee Absenteeism Scale (EAS; Paget, Lang, & Shultz, 1998) was designed to assess employee withdrawal (i.e., not being at work). The EAS is a paper-and-pencil adaptation of Nicholson and Payne's (1987) 12-item open-ended interview. Factor analysis has shown that the items represent two factors: voluntary absences and involuntary absences. The internal

consistency coefficients for these factors are .87 and .67, respectively.

The 81-item Employee Reliability Inventory (Borofsky, 1998, 2000) includes seven scales that assess factors of reliable and productive work behavior (e.g., Emotional Maturity, Freedom from Disruptive Alcohol and Illegal Drug Use, Courtesy, Conscientiousness, Trustworthiness, Long-Term Job Commitment, and Safe Job Performance). Test–retest coefficients over 7 to 21 days ranged from .73 (Trustworthiness) to .89 (Freedom from Disruptive Alcohol and Illegal Drug Use). Scores on the scales differentiate employees who have demonstrated unreliable behaviors for occupationally and geographically diverse job applicant samples. Scale scores predict inventory shrinkage (i.e., missing merchandise) and supervisors' ratings of turnover, unauthorized absence, and overall job performance (Borofsky, 2000). The 22-item On-The-Job Behaviors scale (Lehman & Simpson, 1992) assesses four categories of behavior: psychological withdrawal (e.g., daydreaming), physical withdrawal (e.g., sleeping at work), positive work behaviors (working overtime), and antagonistic work behaviors (arguing). Coefficient alphas range from .84 for psychological withdrawal to .58 for physical withdrawal. Positive Work Behaviors scale scores correlate positively with job satisfaction, job involvement, job tension, and general fatigue. Antagonistic Behaviors scale scores also correlate with job tension and fatigue as well as burnout. Psychological Withdrawal scale scores correlate positively with turnover intentions, general fatigue, and burnout and negatively with perceived organizational support and job satisfaction (Fields, 2002).

Aquino, Lewis, and Bradfield (1999) developed the 14-item Deviant Behaviors measure to assess interpersonal and organizational deviance. The Interpersonal Deviance Scale assesses behaviors, such as making racial slurs and obscene gestures, that inflict harm on others. The Organizational Deviance Scale assesses poor performance such as calling in sick or ignoring instructions. The coefficient alphas are .73 and .76, respectively, for Interpersonal Deviance and Organizational Deviance. Confirmatory factor analysis indicates that the two scales are distinct from each other and distinct from distributive, interactive and procedural justice (Aquino et al., 1999). The nine-item Antisocial Behaviors scale (Robinson & O'Leary-Kelly, 1998) measures behaviors that harm the organization or individuals. Items tap behaviors such as breaking rules, damaging property, being rude, starting arguments, and verbally attacking others. Coefficient alpha values range from .68 to .81. Scores for the Antisocial Behavior scale correlate negatively with organizational citizenship behaviors.

The Organizational Citizenship Behavior Checklist (Fox & Spector, 2009) is a 36-item survey that has two subscales that assesses positive acts directed toward (a) the organization or (b) coworkers. The internal consistencies for the subscales are .92 and .91, respectively.

Interpersonal conflict. Interpersonal conflict at work, in the form of aggression, mistreatment, bullying or harassment of colleagues, subordinates, or superiors, can have a devastating effect on employees and the organization. Uncivil behaviors at work that humiliate, intimidate, frighten, or punish others are common occurrences in the workplace. When such behaviors are directed repeatedly at the same individual over a period of time, the aggregated effect turns into extreme social stress that can cause serious harm. Physically intimidating actions and physical violence or threat of violence are also behaviors that fall under the bullying umbrella. An imbalance of power between the parties is also a component of bullying.

The 22-item Negative Acts Questionnaire–Revised (NAQ-R; Einarsen, Hoele, & Notelaers, 2009), based on the NAQ (Einarsen & Raknes, 1991), was constructed to assess direct and indirect bullying that is work related or person related or is physical intimidation. Cronbach's alpha for the NAQ-R is 0.90. The best fit model for the NAQ-R is the three-factor solution, yet the correlation between the three factors is very high: .83 between person-related and physically intimidating bullying, .96 between person-related and work-related bullying, and .89 between work-related and physically intimidating bullying. This finding suggests a co-occurrence of bullying and that teasing apart the different types of bullying is difficult. The relation between NAQ-R scores and psychosomatic complaints is moderately

strong, and high scores are associated with reduced performance, more health complaints, greater inclination to leave work, and increased sickness-related absenteeism (Einarsen et al., 2009).

The Interpersonal Conflict at Work Scale (ICAWS; Spector & Jex, 1998) was developed to measure conflict with other people at work. The alpha coefficient is .74. The ICAWS scores correlate .41 with intention to quit, .38 with depression, .36 with anxiety, .32 with frustration, and −.32 with job satisfaction. ICAWS scores also correlate .40 with role conflict, .33 with negative affectivity, and .29 with role ambiguity.

The 33-item Workplace Aggression Scale (Rutter & Hine, 2005) assesses the frequency with which individuals exhibit aggressive behaviors. Its two subscales have internal reliability coefficients of .79 (Obstructionism subscale) and .82 (Overt Aggression subscale). Factor analysis supports a three-factor model: behaviors expressing hostility; behaviors designed to interfere with another's performance, and behaviors expressing physical assault (Pseekos, Bullock-Yowell, & Dahlen, 2011).

Stress and strain. Social support and WIF/FIW are two variables that serve multiple roles in OHP stress-strain models. For example, lack of support within an organization may be a job stressor that predicts emotional or behavioral strain. On the other hand, the presence of support, inside or outside the organization, may buffer employees' experiences of work stress and reduce the possibility of strain. WIF or FIW may be stressors that result in work strain or in personal life strain. However, other forms of job stress (e.g., bullying, job demand, shift work) that correlate with many markers of job strain also correlate with WIF, suggesting that WIF may be a strain on some occasions as well as a stressor on other occasions.

Assessment of social support. Social support includes both tangible assistance and feelings of emotional support that serve to have a buffering effect that reduces the negative effects of stress. Measures of social support assess workplace support as well as social support from friends, family, and significant others. Within the workplace, the support may be characterized as organizational support (the extent to which an organization cares about employee well-being) or as support provided by agents of the organizations (e.g., supervisors, coworkers).

Eisenberger, Huntington, Hutchinson, and Sowa (1986) developed the Survey of Perceived Organizational Support (SPOS) to measure an employee's perception that an organization values her or his contributions (eight items) and that the organization engages in activities that promote employee well-being (nine items). They reported a Cronbach's alpha of 0.97. The SPOS scores correlate positively with pay and recognition expectations, overall job satisfaction, and employee organizational commitment and negatively with turnover intentions, perceived organizational politics, emotional exhaustion, and days absent (Fields, 2002). Kottke and Sharafinski (1988) modified the wording of the 16 SPOS items, by changing *organization* to *supervisor*, to develop the Survey of Perceived Supervisory Support (SPSS). They reported a Cronbach's alpha of 0.98 for the scale. Although the correlation of scores for the two scales for a large sample of city employees is only .13 (Kottke & Sharafinski, 1988), the relation of scores on the SPSS to SPOS scores increases with perceived supervisor status in the organization (Eisenberger, Stinglhamber, Vandenberghe, Sucharski, & Rhodes, 2002).

Greenhaus, Parasuraman, and Wormley (1990) developed a nine-item measure of supervisory support that taps the extent to which supervisors are invested in an employees' career development and job performance. They report a coefficient alpha level of .93. Scale scores correlate positively with employee promotions, career satisfaction, and job performance.

A measure of social support, developed by Caplan et al. (1980), includes three subscales that assess support from coworkers and supervisors, significant others, and family and friends. The scales tap both emotional and instrumental support. Coefficient alphas range from .86 to .91. Scale scores correlate negatively with job insecurity and noncompliant job behaviors and positively with job satisfaction and work group cohesiveness (Fields, 2002).

Parasuraman, Greenhaus, and Granrose (1992) developed an eight-item scale to assess spouse/

partner support. Factor analysis identified two factors—emotional support and informational support—that correlated .63 for women and .72 for men; thus, the items were aggregated into one scale. The alpha coefficients are .86 and .91, respectively, for women and men.

WIF/FIW assessment. Research in OHP has focused increasingly on the demands of work and family roles on psychological well-being. Demanding or unrewarding work may lead to increased family strain (i.e., WIF), whereas positive work experiences, on the other hand, may reduce strain. Unsatisfactory relationships and stressful family situations may have a negative effect on psychological reactions to work as well as work behaviors (i.e., FIW). Research shows that, as individuals experience increased conflict between work and family roles, job and life satisfaction deteriorates along with organizational commitment, and disruptive work behaviors increase (e.g., tardiness and absenteeism; Aryee, Luk, & Stone, 1998). Research also provides evidence that, although some variance is shared between WIF and FIW, the constructs are distinct. Thus, WIF or FIW may be the cause or the consequence of psychological strain.

A plethora of scales are used to measure WIF or FIW. The Multidimensional Measure of Work–Family Conflict (Carlson, Kacmar, & Williams, 2000) assesses three forms of conflict (time, strain, and behavior) and the two directions of conflict (WIF and FIW). Coefficient alphas are .87 for time-based WIF, .79 for time-based FIW, .85 for strain-based WIF, .87 for strain-based FIW, .78 for behavior-based WIF, and .85 for behavior-based FIW. As expected, the correlations among WIF subscales and among FIW subscales generally are larger than the correlations between the two sets of subscales. WIF Strain and Behavior conflict scale scores are significantly related to life and family satisfaction. Role conflict and social support predict FIW time- and strain-based conflict. Organizational commitment is significantly related to FIW behavior-based conflict. Matthews, Kath, and Barnes-Farrell (2010) developed two three-item abbreviated versions of Carlson et al.'s measure for use in research that has assessment time constraints.

Another instrument that assesses both WIF and FIW was developed by Gutek, Searle, and Klepa (1991) and augmented by Carlson and Perrewé (1999). The two subscales (each six items) can be combined into a composite measure of work and family interference. The WIF subscale alpha coefficients range from .71 to .87, and the FIW coefficients range from .74 to .83. Coefficient alphas for the composite measure range from .66 to .89 (Fields, 2002; Gutek et al., 1991). Positive correlations with the WIF subscale scores include hours spent at work, flextime, number of children at home, child care needs, and working at home. Many of the same variables correlate positively with FIW scale scores (e.g., child care needs, flextime, working at home) as well as with family involvement and family conflict. Scores on both subscales correlate negatively with job satisfaction, life satisfaction, age of youngest child, and control over hours worked (Fields, 2002). Structural equation models and factor analysis indicate that FIW and WIF covary but nonetheless are empirically distinct (Fields, 2002).

LOOKING AHEAD

Much work in OHP that uses testing and assessment has been research oriented and has focused on testing models and theories of work-related stress. A multitude of instruments have been developed for measuring stressors, strains, and health outcomes to conduct this research,. However, OHP is both a scientific discipline and an applied field. Directions for the field's next period of evolution most likely will include (a) an increased emphasis on workplace interventions, (b) an expanded interest in understanding coping mechanisms, and (c) an increased awareness of cross-cultural issues.

Workplace Interventions

To date, few papers that report intervention studies have been published (Kang, Staniford, Dollard, & Kompier, 2008). However, intervention studies are essential translational research. The most fruitful work will not only examine the outcomes of interventions (e.g., reduced stressors or strains, positive psychological and physical health outcomes) but also will seek to evaluate the intervention process.

Such research designs are especially valuable in the event that an intervention appears to fail to meet the specified outcome criteria. Regardless of the research design (e.g., quasi-experimental, mixed methods) these evaluations will require assessment to understand fully the intervention process and intervention outcomes. In response to evidence that employees vary considerably in the way in which they appraise their experience of an intervention, Randall, Nielsen, and Tvedt (2009) developed five scales that measure employees' evaluations of workplace stress management. The Intervention Process Measure (IPM) includes scales that assess manager attitude toward the intervention, employee reactions to exposure to the intervention, employee involvement in the intervention design and implementation, employee readiness for change, and intervention history. Confirmatory factor analysis indicates acceptable fit for a five-factor model and unacceptable fit for a one-factor model. Correlations between the five scales are modest, and Cronbach's alphas range from 0.65 (the two-item Intervention History scale) to 0.89 (the seven-item Line Manager Attitudes and Actions scale). The implementation of scales such as those of the IPM may prove especially useful in settings that do not allow the use of a control group (e.g., Type III error, when examining the intervention outcome scores, can be reduced by identifying employees with high or low scores on the process variables).

Coping Strategies

One of the knotty problems in OHP assessment is how to measure aspects of coping in stressful situations. Within the workplace, careful assessment of coping strategies is needed (a) to understand the role that coping may play in reducing strain, (b) to design and evaluate interventions to strengthen coping skills, and (c) to understand the interaction between coping and individual differences constructs such as resilience, positive and negative affect, or hardiness. Latack (1986) developed scales to measure the extent to which an individual, in a work setting, uses three coping strategies: control (including actions and cognitive reappraisals that are proactive), escape (actions and cognitive reappraisals that suggests avoidance), and symptom management (strategies to manage symptoms related to job stress). The initial item pool included 23 control and escape action items, 14 cognitive reappraisal items, and 27 symptom management items. Cluster analysis confirmed that the items represented a proactive, take-charge cluster (i.e., control; $n = 17$ items), a mental and behavioral avoidance cluster (i.e., escape; $n = 11$ items) and a symptom management cluster ($n = 24$ items). Coefficient alpha levels are .70 for symptom management, .71 for escape, and .85 for control. Individuals facing personal life changes are most likely to score high on the Symptom Management Coping scale, whereas people facing job-related stressors seem to use all three strategies. High scores on the Control scale correlate positively with social support and job satisfaction and negatively with anxiety and propensity to leave. High scores on the Escape scale and the Symptom Management scale correlate with psychosomatic symptoms and complaints.

Cross-Cultural Issues

The European Academy of Occupational Health Psychology emerged in the late 1990s to bring together professionals engaged in research, teaching, and practice related to psychological, social, and organizational issues in occupational health. In 2004, a parallel organization, the Society for Occupational Health Psychology, emerged in North America. These two organizations now are part of the International Coordinating Group for Occupational Health Psychology, created in 2000 to encourage international collaboration (Houdmont & Leka, 2010). This internationalization of the OHP movement provides the basis for a science and practice that can incorporate underlying societal culture and values as important components in research and practice. Economic globalization and the increased rate of immigration in many countries make an understanding of cultural and ethnic perceptions of stressors and strains an important focus in OHP research and practice. The need to understand the effect of culture is critical to testing and assessment both in translation of instruments for use in many languages as well as in the understanding and interpretation of the constructs the instruments assess (Riordan & Vandenberg, 1994). The use of translated versions of tests and

questionnaires requires attention to conceptual equivalence, functional equivalence, and measurement equivalence. Because literal translations may be interpreted differently across languages and across cultures, care should be taken to ensure equivalence of meaning as well as to ensure that attitudinal items are culturally appropriate (Glazer, 2008).

CONCLUSION

The field of OHP relies heavily on testing and assessment to pursue both a scientific research agenda as well as applied practice. Critical to this work is attention to developing and selecting tests, questionnaires, and assessments with established evidence of reliability and validity. One neglected area for most OHP scale development is the investigation of instrument test–retest reliability. Providing evidence of stability is essential to ongoing efforts to understand whether stress is a state or trait of the individual and the work environment. As OHP establishes a practice foothold in organizations, the accumulation of evidence of construct, criterion and content validity for job stress and strain measures also is essential especially in light of job-stress-related litigation. As the workforce evolves with the increased employment of women, ethnic minorities, and older workers, existing OHP measures need to be scrutinized to determine whether they generalize to emerging segments of workers.

References

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology, 63*, 1–18. doi:10.1111/j.2044-8325.1990.tb00506.x

Aquino, K., Lewis, M. U., & Bradfield, M. (1999). Justice constructs, negative affectivity, and employee deviance: A proposed model and empirical test. *Journal of Organizational Behavior, 20*, 1073–1091. doi:10.1002/(SICI)1099-1379(199912)20:7<1073::AID-JOB943>3.0.CO;2-7

Aryee, S., Luk, V., & Stone, R. (1998). Family responsive variables and retention-relevant outcomes among employed parents. *Human Relations, 51*, 73–87. doi:10.1177/001872679805100105

Bacharach, S. B., Bamberger, P., & Conley, S. C. (1990). Work processes, role conflict, and role overload: The case of nurses and engineers in the public sector. *Work and Occupations, 17*, 199–228. doi:10.1177/0730888490017002004

Bachiochi, P. D., & Weiner, S. P. (2002). Qualitative data collection and analysis. In S. G. Rogelberg (Ed.), *Handbook of organizational research methods* (pp. 161–183). Malden, MA: Wiley-Blackwell.

Bakker, A. B., Demerouti, E., & Schaufeli, W. B. (2002). Validation of the Maslach Burnout Inventory-General Survey: An Internet study across occupations. *Anxiety, Stress, and Coping, 15*, 245–260. doi:10.1080/1061580021000020716

Beehr, T. A., Walsh, J. T., & Taber, T. D. (1976). Relationship of stress to individually and organizationally valued states: Higher order needs as a moderator. *Journal of Applied Psychology, 61*, 41–47. doi:10.1037/0021-9010.61.1.41

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*, 579–616. doi:10.1146/annurev.psych.54.101601.145030

Borofsky, G. L. (1998). *User's manual for the Employee Reliability Inventory*. Boston, MA: Bay State Psychological Association.

Borofsky, G. L. (2000). Predicting involuntary dismissal for unauthorized absence, lateness, and poor performance in the selection of unskilled and semiskilled British contract factory operatives: The contribution of the Employee Reliability Inventory. *Psychological Reports, 87*, 95–104.

Breaugh, J. A., & Colihan, J. P. (1994). Measuring facets of job ambiguity: Construct validity evidence. *Journal of Applied Psychology, 79*, 191–202. doi:10.1037/0021-9010.79.2.191

Caplan, R. D., Cobb, S., French, J. R. P., Van Harrison, R., & Pinneau, S. R. (1980). *Job demands and worker health*. Ann Arbor: University of Michigan, Institute for Social Research.

Carlopio, J. R. (1996). Construct validity of a Physical Work Environment Satisfaction Questionnaire. *Journal of Occupational Health Psychology, 1*, 330–344. doi:10.1037/1076-8998.1.3.330

Carlson, D. S., Kacmar, K. M., & Williams, L. J. (2000). Construction and initial validation of a multidimensional measure of work-family conflict. *Journal of Vocational Behavior, 56*, 249–276. doi:10.1006/jvbe.1999.1713

Carlson, D. S., & Perrewé, P. L. (1999). The role of social support in stressor-strain relationship: An examination of work-family conflict. *Journal of Management, 25*, 513–540. doi:10.1177/014920639902500403

Chandola, T., Heraclides, A., & Kumari, M. (2010). Psychophysiological biomarkers of workplace stressors. *Neuroscience and Biobehavioral Reviews, 35*, 51–57. doi:10.1016/j.neubiorev.2009.11.005

Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80*, 265–278. doi:10.1016/j.biopsycho.2008.10.004

Cohen, A., & Margolis, B. (1973). Initial psychological research related to the Occupational Safety and Health Act of 1970. *American Psychologist*, *28*, 600–606. doi:10.1037/h0034997

Connell, P., Lee, B. V., & Spector, P. E. (2004). Job stress assessment methods. In J. C. Thomas & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment* (pp. 455–469). Hoboken, NJ: Wiley.

Cooper, C. L., Dewe, P. J., & O'Driscoll, M. P. (2001). *Organizational stress: A review and critique of theory, research, and applications*. Thousand Oaks, CA: Sage.

Demerouti, E., & Bakker, A. B. (2008). The Oldenburg Burnout Inventory: A good alternative to measure burnout (and engagement). In J. R. B. Halbesleben (Ed.), *Handbook of stress and burnout in health care* (pp. 65–78). New York, NY: Nova Science Publishers.

Demerouti, E., Bakker, A. B., Vardakou, I., & Kantas, A. (2003). The convergent validity of two burnout instruments: A multitrait–multimethod analysis. *European Journal of Psychological Assessment, 18*, 296–307.

Dwyer, D. J., & Ganster, D. C. (1991). The effects of job demands and control in employee attendance and satisfaction. *Journal of Organizational Behavior, 12*, 595–608. doi:10.1002/job.4030120704

Eggerth, D. E., & Cunningham, T. R. (2012). Counseling psychology and occupational health psychology. In E. Altamaier & J. C. Hansen (Eds), *Handbook of counseling psychology* (pp. 752–779). New York, NY: Oxford University Press.

Einarsen, S., Hoele, H., & Notelaers, G. (2009). Measuring exposure to bullying and harassment at work: Validity, factor structure and psychometric properties of the Negative Acts Questionnaire—Revised. *Work and Stress, 23*, 24–44. doi:10.1080/02678370902815673

Einarsen, S., & Raknes, B. (1997). Harassment in the workplace and victimization of men. *Violence and Victims, 12*, 247–263.

Eisenberger, R., Huntington, R., Hutchinson, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology, 71*, 500–507. doi:10.1037/0021-9010.71.3.500

Eisenberger, R., Stinglhamber, F., Vandenberghe, C. V., Sucharski, I. L., & Rhodes, L. (2002). Perceived supervisor support: Contributions to perceived organizational support and employee retention. *Journal of Applied Psychology, 87*, 565–573. doi:10.1037/0021-9010.87.3.565

Fields, D. L. (2002). *Taking the measure of work*. Thousand Oaks, CA: Sage.

Fox, S., & Spector, P. E. (2009). *Organizational Citizenship Behavior Checklist (OCB-C)*. Unpublished manuscript, University of South Florida, Tampa. Retrieved from http://shell.cas.usf.edu/pspector/scales/OCB-C development.doc

Fried, Y., Rowland, K. M., & Ferris, G. R. (1984). The physiological measurement of work stress: A critique. *Personnel Psychology, 37*, 583–615. doi:10.1111/j.1744-6570.1984.tb00528.x

Ganster, D. C. (1989). Worker control and well-being: A review of research in the workplace. In S. Sauter, J. Hurrell, & C. Cooper (Eds.), *Job control and worker health* (pp. 3–24). Chichester, England: Wiley.

Glazer, S. (2008). Cross-cultural issues in stress and burnout. In J. R. B. Halbesleben (Ed.). *Handbook of stress and burnout in health care* (pp. 79–93). New York, NY: Nove Science Publishers.

Goetzel, R. Z., Anderson, D. R., Whitmer, R. W., Ozminkowski, R. J., Dunn, R. L., & Wasserman, J. (1998). The relationship between modifiable health risks and health care expenditures: An analysis of the multi-employer HERO health risk and cost database. *Journal of Occupational and Environmental Medicine, 40*, 843–854. doi:10.1097/00043764-199810000-00003

Gray, M. P., & O'Brian, K. M. (2007). Addressing the assessment of women's career aspirations: The Career Aspiration Scale. *Journal of Career Assessment, 15*, 317–337. doi:10.1177/1069072707301211

Greenhaus, J. H., Parasuraman, A., & Wormley, W. M. (1990). Effects of race on organizational experience, job performance evaluations, and career outcomes. *Academy of Management Journal, 33*, 64–86. doi:10.2307/256352

Gutek, B. A., Searle, S., & Klepa, L. (1991). Rational versus gender role explanations for work–family conflict. *Journal of Applied Psychology, 76*, 560–568. doi:10.1037/0021-9010.76.4.560

Hackman, J. R., & Oldman, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*, 159–170. doi:10.1037/h0076546

Halbesleben, J. R. B., & Demerouti, E. (2005). The construct validity of an alternative measure of burnout: Investigating the English translation of the Oldenburg Burnout Inventory. *Work and Stress, 19*, 208–220. doi:10.1080/02678370500340728

Hansen, A. M., Larsen, A. D., Rugulies, R., Garde, A. H., & Knudsen, L. E. (2009). A review of the effect of the psychosocial working environment on physiological changes in blood and urine. *Basic and Clinical Pharmacology and Toxicology, 105*, 73–83. doi:10.1111/j.1742-7843.2009.00444.x

Hartman, D. P., Barios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In S. N. Haynes

& E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment* (pp. 108–127). Hoboken, NJ: Wiley.

Hodson, R., Creighton, S., Jamison, C. S., Rieble, S., & Welsh, S. (1994). Loyalty to whom? Workplace participation and the development of consent. *Human Relations, 47*, 895–909. doi:10.1177/001872679404700802

Houdmont, J., & Leka, S. (2010). An introduction to occupational health psychology. In S. Leka & J. Houdmont (Eds.), *Occupational health psychology* (pp. 1–30). Malden, MA: Wiley-Blackwell.

House, R. J., Schuler, R. S., & Levanoni, E. (1983).). Role conflict and ambiguity scales: Reality or artifacts? *Journal of Applied Psychology, 68*, 334–337. doi:10.1037/0021-9010.68.2.334

Hurrell, J. J., Jr., & McLaney, M. A. (1988). Exposure to job stress: A new psychometric instrument. *Scandinavian Journal of Work, Environment, and Health, 14*, 27–28.

Hurrell, J. J., Jr., Nelson, D. L., & Simmons, B. L. (1998). Measuring job stressors and strains: Where we have been, where we are, and where we need to go. *Journal of Occupational Health Psychology, 3*, 368–389. doi:10.1037/1076-8998.3.4.368

Ironson, G. H., Smith, P. C., Brannick, M. T., Gibson, W. M., & Paul, K. B. (1989). Constitution of a Job in General scale: A comparison of global composite and specific measures. *Journal of Applied Psychology, 74*, 193–200. doi:10.1037/0021-9010.74.2.193

Kang, S. Y., Staniford, A. K., Dollard, M. F., & Kompier, M. A. J. (2008). Knowledge development and content in occupational health psychology: A systematic analysis of the *Journal of Occupational Health Psychology* and *Work and Stress*, 1996–2006. In J. Houdmont & S. Leka (Eds.), *Occupational health psychology: European perspectives on research, education and practice* (Vol. 3, pp. 27–63). Maia, Portugal: ISMAI Publishers.

Karasek, R. A. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly, 24*, 285–308. doi:10.2307/2392498

Kessler, R. C., Barber, C., Beck, A., Berglund, P., Cleary, P. D., McKenas, D., . . . Wang, P. (2003). The World Health Organization Health and Work Performance Questionnaire (HPQ). *Journal of Occupational and Environmental Medicine*, *45*, 156–174. doi:10.1097/01.jom.0000052967.43131.51

Kiecolt-Glaser, J. K., McGuire, L., Robles, T. F., & Glaser, R. (2002). Emotions, morbidity, and mortality. *Annual Review of Psychology, 53*, 83–107. doi:10.1146/annurev.psych.53.100901.135217

Kottke, J. L., & Sharafinski, C. E. (1988). Measuring perceived supervisory and organizational support. *Educational and Psychological Measurement, 48*, 1075–1079. doi:10.1177/0013164488484024

Krantz, D. S., & McCeney, M. K. (2002). Effects of psychological and social factors on organic disease: A critical assessment of research on coronary heart disease. *Annual Review of Psychology, 53*, 341–369. doi:10.1146/annurev.psych.53.100901.135208

Latack, J. (1986). Coping with job stress: Measures and future directions for scale development. *Journal of Applied Psychology, 71*, 377–385. doi:10.1037/0021-9010.71.3.377

Lehman, W. E. K., & Simpson, D. D. (1992). Employee substance use and on-the-job behaviors. *Journal of Applied Psychology, 77*, 309–321. doi:10.1037/0021-9010.77.3.309

Leka, S., & Cox, T. (2010). Psychosocial risk management at the workplace level. In S. Leka & J. Houdmont (Eds.), *Occupational health psychology* (pp. 124–156). Maldon, MA: Wiley-Blackwell.

Marrow, P. (1993). *The theory and measurement of work commitment*. Greenwich, CT: JAI Press.

Maslach, C., & Jackson, S. E. (1986). *Maslach Burnout Inventory manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Maslach, C., Jackson, S. E., & Leiter, M. (1996). *Maslach Burnout Inventory manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Mathieu, J. E., & Zajac, D. M. (1990). A review and meta-analysis of the antecedents, correlates, and consequences of organizational commitment. *Psychological Bulletin*, *108*, 171–194. doi:10.1037/0033-2909.108.2.171

Matthews, R. A., Kath, L. M., & Barnes-Farrell, J. L. (2010). A short, valid, predictive measure of work-family conflict: Item selection and scale validation. *Journal of Occupational Health Psychology, 15*, 75–90. doi:10.1037/a0017443

Melamed, S., Ben-Avi, I., Luz, J., & Green, M. S. (1995). Objective and subjective work monotony: Effects on job satisfaction, psychological stress, and absenteeism in blue-collar workers. *Journal of Applied Psychology, 80*, 29–42. doi:10.1037/0021-9010.80.1.29

Meyer, J. P. (1997). Organizational commitment. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 175–228). New York, NY: Wiley.

Meyer, J. P., & Allen, N. J. (1997). *Commitment in the workplace.* Thousand Oaks, CA: Sage.

Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior, 14*, 224–247. doi:10.1016/0001-8791(79)90072-1

Nakata, A., Takahashi, M., Irie, M., & Swanson, N. G. (2010). Job satisfaction as associated with elevated natural killer cell immunity among healthy white-collar employees. *Brain, Behavior, and Immunity, 24*, 1268–1275. doi:10.1016/j.bbi.2010.05.004

Nakata, A., Takahashi, M., Otsuka, Y., & Swanson, N. G. (2010). Is self-rated health associated with blood immune markers in healthy individuals? *International Journal of Behavioral Medicine, 17*, 234–242. doi:10.1007/s12529-010-9102-0

National Institute of Occupational Safety and Health. (2004). *Stress at work.* Retrieved from http://www.cdc.gov/niosh/ohp.html

Netemeyer, R. G., Johnston, M. W., & Burton, S. (1990). Analysis of role conflict and role ambiguity in a structural equations framework. *Journal of Applied Psychology, 75*, 148–157. doi:10.1037/0021-9010.75.2.148

Nicholson, N., & Payne, R. (1987). Absence from work: Explanations and attributions. *Applied Psychology, 36*, 121–132. doi:10.1111/j.1464-0597.1987.tb00379.x

Niehoff, B. P., & Moorman, R. H. (1993). Justice as a mediator of the relationship between methods of monitoring and organizational citizenship. *Academy of Management Journal, 36*, 527–556. doi:10.2307/256591

O'Brien, K. M. (1996). The influence of psychological separation and parental attachment on the career development of adolescent women. *Journal of Vocational Behavior, 48*, 257–274. doi:10.1006/jvbe.1996.0024

Occupational Safety and Health Act of 1970. P.L. 91–596. 91st Congress, S. 2193, December 29, 1970.

O'Driscoll, M. P., & Brough, P. (2010). Work organization and health. In S. Leka & J. Houdmont (Eds.), *Occupational health psychology* (pp. 57–87). Malden, MA: Wiley-Blackwell.

Paget, K. J., Lang, D. L., & Shultz, K. S. (1998). Development and validation of an employee absenteeism scale. *Psychological Reports, 82*, 1144–1146.

Parasuraman, S., Greenhaus, J., & Granrose, C. (1992). Role stressors, social support and well-being among two career couples. *Journal of Organizational Behavior, 13*, 339–356. doi:10.1002/job.4030130403

Pines, A. M., & Aronson, E. (1988). *Career burnout: Cause and cures* (2nd ed.). New York, NY: Free Press.

Prasad, M., Wahlquist, P., Shikiar, R., & Shih, Y.-C. T. (2004). A review of self-report instruments measuring health-related work productivity: A patient-report outcomes perspective. *PharmacoEconomics, 22*, 225–244. doi:10.2165/00019053-200422040-00002

Pseekos, A. C., Bullock-Yowell, E., & Dahlen, E. (2011). Examining Holland's person-environment fit, workplace aggression, interpersonal conflict, and job satisfaction. *Journal of Employment Counseling, 48*, 63–71. doi:10.1002/j.2161-1920.2011.tb00115.x

Randall, R., Nielsen, K., & Tvedt, S. D. (2009). The development of five scales to measure employees' appraisals of organizational-level stress management interventions. *Work and Stress, 23*, 1–23. doi:10.1080/02678370902815277

Reilly, M. C., Zbrozek, A. S., & Dukes, E. M. (1993). The validity and reproducibility of a work productivity and activity impairment instrument. *PharmacoEconomics, 4*, 353–365. doi:10.2165/00019053-199304050-00006

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*, 643–671.

Rizzo, J., House, R. J., & Lirtzman, S. I. (1970). Role conflict and role ambiguity in complex organizations. *Administrative Science Quarterly, 15*, 150–163. doi:10.2307/2391486

Robinson, S. L., & O'Leary-Kelly, A. M. (1998). Monkey see, monkey do: The influence of work groups on the antisocial behavior of employees. *Academy of Management Journal, 41*, 658–672. doi:10.2307/256963

Roznowski, M. (1989). An examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology, 74*, 805–814. doi:10.1037/0021-9010.74.5.805

Rutter, A., & Hine, D. W. (2005). Sex differences in workplace aggression: An investigation of moderation and mediation effects. *Aggressive Behavior, 31*, 254–270. doi:10.1002/ab.20051

Sauter, S. L., Hurrell, J. J., Jr., Fox, H. R., Tetrick, L. E., & Barling, J. (1999). Occupational health psychology: An emerging discipline. *Industrial Health, 37*, 199–211. doi:10.2486/indhealth.37.199

Shirom, A. (2010). Employee burnout and health. In J. Houdmont & S. Leka (Eds.), *Contemporary occupational health psychology* (pp. 59–76). Malden, MA: Wiley-Blackwell.

Shore, L. M., & Tetrick, L. E. (1991). A construct validity study of the Survey of Perceived Organizational Support. *Journal of Applied Psychology, 76*, 637–643. doi:10.1037/0021-9010.76.5.637

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *Measurement of satisfaction in work and retirement*. Chicago, IL: Rand McNally.

Sonnentag, S., & Fritz, C. (2006). Endocrinological processes associated with job stress: Catecholamine and cortisol responses to acute and chronic stressors. In P. L. Perrewé & D. C. Ganster (Eds.), *Employee health, coping and methodologies* (pp. 1–59). San Francisco, CA: Elsevier.

Spector, P. E., & Fox, S. (2003). Reducing subjectivity in the assessment of the job environment: Development

of the Factual Autonomy Scale (FAS). *Journal of Organizational Behavior, 24*, 417–432. doi:10.1002/job.199

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain. *Journal of Occupational Health Psychology, 3*, 356–367. doi:10.1037/1076-8998.3.4.356

Spielberger, C. D. (1994). *Professional manual for the Job Stress Survey (JSS)*. Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D., & Reheiser, E. C. (1995). Measuring occupational stress: The Job Stress Survey. In R. Grandal & P. L. Perrewe (Eds.), *Occupational Stress* (pp. 51–69). New York, NY: Taylor & Francis.

Stewart, W. F., Lipton, R. B., Kolodner, K., Liberman, J., & Sawyer, J. (1999). Reliability of the migraine disability assessment score in a population-based sample of headache sufferers. *Cephalalgia, 19*, 107–114. doi:10.1046/j.1468-2982.1999.019002107.x

Sweeney, P. D., & McFarlin, D. B. (1997). Process and outcome: Gender differences in the assessment of justice. *Journal of Organizational Behavior, 18*, 83–98. doi:10.1002/(SICI)1099-1379(199701)18:1<83::AID-JOB779>3.0.CO;2-3

Taris, T. W., deLange, A. H., & Kompier, M. A. J. (2010). Research methods in occupational health psychology. In S. Leka & J. Houdmont (Eds.), *Occupational health psychology* (pp. 269–297). Malden, MA: Wiley-Blackwell.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey responses*. New York, NY: Cambridge University Press.

Tryon, W. W. (1998). Behavioral observation: In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 70–103). Boston, MA: Allyn & Bacon.

Van Katwyk, P. T., Fox, S., Spector, P. E., & Kelloway, E. K. (2000). Using the Job-Related Affective Well-Being Scale (JAWS) to investigate affective responses to work stressors. *Journal of Occupational Health Psychology, 5*, 219–230. doi:10.1037/1076-8998.5.2.219

Van Veldhoven, M., Broeresen, J., & Fortuin, R. (1999). *The picture of job stress. Psychosocial work load and job stress in the Netherlands*. Amsterdam, the Netherlands: Stichtina.

Van Veldhoven, M., & Meijman (1994). *The measurement of psychosocial demands*. Amsterdam, the Netherlands: Nederlands Instituut voor Arbeidsom Standigheden.

Van Yperen, N. W., & Snijders, T. A. P. (2000). A multilevel analysis of the demands-control model: Is stress at work determined by factors at the group level or the individual level? *Journal of Occupational Health Psychology, 5*, 182–190. doi:10.1037/1076-8998.5.1.182

Weiss, D. J., Dawis, R., Lofquist, L. H., & England, G. W. (1966). *Minnesota Studies in Vocational Rehabilitation: Vol. 21. Instrumentation for the theory of work adjustment*. Minneapolis: University of Minnesota Press.

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59*, 339–354. doi:10.1111/j.1467-6494.1991.tb00252.x

Work in America. (1974). *A report of a special task force to the Secretary of Health Education and Welfare*. Cambridge, MA: MIT Press.

CHAPTER 31

ASSESSMENT IN SPORT AND EXERCISE PSYCHOLOGY

Todd J. Wilkinson

The chapter begins with a brief introduction to the field of sport and exercise psychology and the content of which it traditionally encompasses, including a brief historical review of the field as it relates to general assessment. This overview provides a context for a discussion of the development and use of research methodologies in sport psychology from its emergence to present day trends and applications. Although sport and exercise psychology assessment involves the use of measures developed outside the context of sport and exercise, the chapter primarily focuses its review on individual assessment instruments specific to the sport and exercise psychology domain. There is an abundance of measures in the field of varying psychometric quality, and this review favors established measures in sport psychology used in both basic and applied research. The review includes evaluation of the psychometric properties of the instruments as well as considerations for their intended use. The chapter concludes with identification and discussion of important issues with the traditional and current research practices, including current needs and directions for future development of testing and assessment in sport psychology.

The field known as sport psychology has traditionally involved the scientific investigation of human behavior in relation to sport but, for some time now, has been broadened to encompass the domain of exercise psychology also. Defined by Division 47 of the American Psychological Association (APA), Sport and Exercise Psychology is "the scientific study of the psychological factors that are associated with participation and performance in sport, exercise, and other types of physical activity" (APA, 2010). This integrated subdiscipline of sport and exercise psychology encompasses a wide range of related constructs, including aggression, arousal and anxiety, attention, cohesion, motivation, confidence, exercise and anxiety, variables related to sport fans, and exercise participation among others.

Although assessment is used generally for the purposes of description, explanation, prediction, and perhaps control of behavior (Cozby, 2007), assessment in sport and exercise psychology has been used in a number of specific areas such as evaluation of psychological skills, performance enhancement, prediction of outcomes and selection, group processes, rehabilitation from injury, clinical issues (e.g., depression, eating disorders) and health and exercise. Assessment has primarily involved the use of psychological inventories, survey questionnaires, behavioral observation, interviews, and psychophysiological measures (Vealey & Garner-Holman, 1998). Assessment of sport-related constructs are reflected in the history of sport psychology.

HISTORY

Sport psychology has its roots in ancient Asian and Greek cultures (Mahoney, 1989). Indeed, Socrates spoke of the psychology of sport of the ancient Greeks and the proper training of athletes (Lavallee, Kremer, Moran, & Williams, 2004). In modern times, sport psychology first took a foothold in the sport and exercise sciences and can be traced back

DOI: 10.1037/14048-031
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

to the 1800s with Norman Triplett's pioneering study of social facilitation in sport, widely cited as the first study in both social and sport psychology (Triplett, 1898). Observing that cyclists expended more effort in the presence of other cyclists than when alone, Triplett designed a similar study in a different sporting context. Triplett found the same social facilitation effect in his experiment that examined the rate at which participants wound the reel of a fishing rod.

Other notable contributors include Coleman Griffith, widely considered the father of sport psychology in North America, who in 1925 established the first American laboratory dedicated to the study of sport psychology-related variables and conducted extensive research at the University of Illinois. Griffith was also hired by the Chicago Cubs in 1938 to provide consulting services and published books and many articles on topics in motor behavior. Griffith was known to conduct his work both within the laboratory as well as on the playing field; indeed, he interviewed the great running back Red Grange after a stellar college performance (scoring four touchdowns in the game's first 12 minutes) to gain insight into possible psychological antecedents and perhaps as a precursor to the study of flow, or optimal experience (Mahoney, 1989). Ahead of her time, Dorothy Yates, also during the first half of the century, trained boxers and aviators in psychological interventions and taught some of the earliest related courses at San Jose State College (Williams & Straub, 2009).

About the same time in Europe, others also were making notable inroads into the field. Carl Diem and R. W. Schulte in Germany, as well as Peter Roudik and A. C. Puni in the Soviet Union (the latter of whom established the Institutes for Physical Culture in the 1920s), were developing sport psychology in their respective countries. The European scientists took a much more applied approach than did the American researchers, but their efforts did not have influence across the ocean. These notable contributions notwithstanding, the record of sport psychology research remained spotty through the first half of the 20th century, typically dominated by studies of motor learning using assessment measures of motor behavior.

The 1950s began a period that Landers (1995) termed "The Formative Years" of sport psychology (which would last through the 1970s), in which the foundation of the field today was established. In the 1950s, the study of sport psychology was beginning to gain more interest, with particular attention given to topics in areas including motor learning and behavior, personality, exercise and anxiety, and motivation, among others. However, it was not until the 1960s that sport psychology emerged as a subdiscipline of psychology and became formally recognized by a broader academic community. The International Society of Sport Psychology hosted the first International Congress of Sport Psychology in Rome, Italy, in 1965. This was followed by the establishment of North American professional organizations in sport psychology such as the American Association of Health, Physical Education, and Recreation and the North American Society for the Psychology of Sport and Physical Activity. It also saw the expansion of research, including in the 1970s the development of two journals specific to sport psychology, the *International Journal of Sport Psychology* and the *Journal of Sport Psychology*. Numerous books, conference proceedings, and other academic endeavors, as well as the formulation of academic programs, were also taking root at this time.

As a subdiscipline of psychology, sport psychology has been influenced by the movements and progress in the discipline, which often has been mirrored by the use of assessment practices in the subdiscipline. A significant portion of the work during this time involved the trait theory in personality, typically the use of clinical measures to assess athlete's personality and sport-related attributes. Following this approach, researchers in sport psychology often took an individual differences perspective focusing on identification of an athlete's traits for evaluation or selection purposes. One controversial example of this was Ogilvie and Tutko's book, *Problem Athletes and How to Handle Them*, published in 1966. The book served to give attention to the developing field of sport psychology, but much of the authors' work was based on their personality inventory, The Athlete Motivation Inventory (AMI). Claims that using the AMI, which was purported as a measure of sport-relevant attributes

that could be used to help inform the selection of high-performing athletes, were unsubstantiated. Without demonstrated evidence of validity, the instrument and its use drew sharp criticism from others in the field. Eventually the scores of the AMI were shown to have limited predictive validity. (It should be noted, however, that Ogilvie made other significant contributions to practice and is now considered by many to be the father of applied sport psychology in North America.)

Meanwhile, the research of others in sport and exercise psychology relied significantly on more established psychological measures of personality traits such as the Cattell Sixteen Personality Factor Questionnaire, the Minnesota Multiphasic Personality Inventory, and the Eysenck Personality Inventory. Despite the use of instruments with significant evidence of reliability and validity, the use of such clinical psychological measures for selection purposes or prediction of behavior in terms of athletic performance or other sport variables was found to be lacking. Few lasting contributions remain from these published studies, although some argue that a more thorough investigation of these articles using modern statistical techniques (e.g., meta-analysis) may reveal more significant findings (Landers, 1995).

Sport and exercise psychology continued to be influenced by movements in broader psychology, and a shifting paradigm during the late 1960s and 1970s that moved psychology's focus from a trait-based to a situationist-based perspective—influenced greatly by the work of Mischel (1968)—was also reflected in sport psychology. Assessment methods and measures attempted to capture the interactionism paradigm, and greater attention was given to the situational context and environmental factors in terms of sport performance and behavior.

The mid-1970s in sport psychology also began to see a movement away from traditional laboratory research and the use of general or clinical psychological measures to place a greater emphasis on sport-specific assessment and the measurement of sport-related variables in the field. Along with the changing psychological perspectives of the time, this movement was initiated to a significant extent in sport psychology by Rainer Martens, who published the Sport Competition Anxiety Test (SCAT) in 1977. The SCAT was one of the first sport-specific measures widely used in the field and provided an example for the development of future sport-specific measures. Martens furthered spurred the subdiscipline in this direction through his famous address at the 1978 Canadian Society for Psychomotor Learning and Sport Psychology (CSPLSP) Conference and subsequent publication, *About Smocks and Jocks* (Martens, 1979), in which he criticized the field for laboratory research that, he purported, did not resemble the real-world experience of the athlete. Martens claimed that this research in artificial settings too often proved of little applied value to sport and urged researchers to develop assessment tools that captured the experience of the athlete and to conduct more field research.

What followed in the 1980s was an emergence of sport-specific measures aimed at assessing the behavior of single participants in an athletic context. These measures reflected a more ideographic approach to research that assessed multiple variables rather than the isolation of few variables as in the controlled research experiments. A strong applied movement in the field was thus underway and a new professional organization, the Association for the Advancement of Applied Sport Psychology, was formed in 1985 (later shortened to the Association for Applied Sport Psychology). The proliferation of new sport-specific measures and research studies was accompanied by the establishment of two applied journals in the 1980s that also reflected movement into more applied work, *The Sport Psychologist* and the *Journal of Applied Sport Psychology*.

At the same time, the social–cognitive approach that rose to prominence in psychology in the 1970s and 1980s also was gaining influence in the field of sport psychology. Researchers began investigating the effects of cognitions, self-perceptions, and situational contexts on factors such as sport performance and participant development. The 1980s also saw an expansion from psychology and sport science to exercise. Many of the new measures created during the 1980s, and those to follow, have been related to exercise psychology.

The 1970s and 1980s served to establish a research base for the field; unfortunately, many of the research studies being published too often relied

less on scientific approaches than on developing applied work. Many researchers appeared to misconstrue the meaning of Martens's words, and the result was less than rigorous investigations with little regard for scientific methods in an effort to find "what works." Although these studies included a plethora of new measures, many of the accompanying tests used in these studies were developed ad hoc for the purposes of a single study, often with little attention to psychometric properties. Indeed, Marsh (1998) claimed that during this yet-developing period of the subdiscipline, it was "evident that the quality of assessments in early sport/exercise work was weak" (p. iv). This trend continued into the 1990s, when research witnessed an expansion of measures developed for the single study, some—but not many—with evidence of validity, and typically without release of such information for public consideration (e.g., Fogarty, 1995; McCann, Jowdy, & Van Raalte, 2002).

With the abundance of new psychological assessment measures in sport psychology, the need for an organizational means for these tools became clear. In 1990, Ostrow published *The Directory of Psychological Tests in the Sport and Exercise Sciences*. This directory contained 175 existing sport psychology assessment instruments, including their purpose, description, construction, psychometric data, availability, and source (a second edition in 1996 followed, which contained 314 instruments and illustrated the further expansion of measures in the field, many of which came from exercise psychology). Not only did the reference serve as an organizational tool for researchers, practitioners and others in the field, but it also provided impetus for the improvement and greater validation processes of these measures. Although not an evaluation of these measures, the directory served to illustrate the variability in quality and documented psychometric properties of these measures.

This sentiment for greater validation of assessment measures was shared by many leading researchers in the field in the 1990s (e.g., Gill, 2000). What followed was a stronger emphasis on the development of sound, sport-specific assessment measures with more rigorous construction and empirical validation processes. The improved state of assessment in the field was illustrated by the publication of *Advances in Sport and Exercise Psychology Measurement* (Duda, 1998). This volume presented a selection of assessment instruments in sport and exercise psychology across a multitude of sport-psychological constructs. Unlike the directory of tests, the publication did evaluate measures and served to highlight examples of those of both sound construction and also those of questionable validity.

The progress made in assessments led Marsh (1998) near the end of the millennium to state, "Thankfully, the hey day [sic] of the 'one shot' instrument seems to have ended" (p. xv). The field saw a significant increase in evidence of measurement with greater construct validation procedures, use of sophisticated and appropriate statistical methods for validation and establishment of norms (e.g., Duda, 1998). Examples include the greater use of confirmatory factor analysis in the validation process rather than reliance primarily on exploratory factor analysis. Other statistical techniques, such as structural equation modeling, have allowed greater multivariate analysis that better reflects the complex nature of sport (e.g., Fung, Ng, & Cheung, 2001). Attention to cultural influences and efforts to gather evidence of cross-cultural validity of instruments (e.g., Isogai et al., 2001) also increased.

REVIEW OF MEASURES

The state of assessment in sport and exercise psychology has improved significantly in the past few decades. Although only one component of the assessment process, existing psychological inventories have demonstrated greater evidence of psychometric properties, and new measures are more often developed with a more comprehensive process of validation using multiple multivariate approaches (e.g., Smith, Smoll, Cumming, & Grossbard, 2006). Although the majority of these assessment measures were developed for research purposes, surveys of sport psychology consultants have revealed use of questionnaires and inventories by about two thirds of consultants in their practice with athletes (e.g., O'Connor, 2004). The selection of appropriate measures is vital to quality research and practice. Among the most widely used sport-specific measures in

research and practice are the Competitive State Anxiety Inventory—2 (CSAI–2), the Sport Anxiety Scale (SAS), the Task and Ego Orientation in Sport Questionnaire (TEOSQ), the Group Environment Questionnaire (GEQ), and the Athletic Skills Coping Inventory—28 (ACSI–28). Despite their widespread use and usefulness, these measures themselves are not without limitations. These measures are presented and described in more detail in the following text. Although not an extensive evaluation of each measure, these reviews serve to illustrate the development of improving measures in the field. The most recent comprehensive directory of sport and exercise psychology measures can be found in Ostrow (1996) and readers interested in a more in-depth review of these and other sport and exercise measures are encouraged to explore Duda (1998).

The CSAI–2

The CSAI–2 (Martens, Vealey, Burton, Bump, & Smith, 1990) is perhaps the most extensively used inventory to measure an athlete's state anxiety related to sport competition. This 27-item measure improved on previous unidimensional measures of anxiety, capturing both the somatic and cognitive components of the construct. The measure contains three subscales (nine items each) measuring somatic anxiety, cognitive anxiety, and self-confidence. Response options use a Likert-type scale ranging from 1 (*not at all*) to 4 (*very much so*), with questions relating to bodily (somatic anxiety) and mental (cognitive anxiety). The inventory has demonstrated adequate internal consistency reliabilities, from .81 to .88 in a validation sample showed to be similar to other samples (Martens et al., 1990). Scores on the CSAI–2 have correlated with other established measures of anxiety, such as the SCAT.

Despite documented evidence for these types of reliability and validity, further analysis of the measure indicated problems with its factor structure (Lane et al., 1999). Using confirmatory factor analysis, Lane et al. (1999) did not replicate the three scales (measuring cognitive anxiety, somatic anxiety, and self-confidence) adequately. Furthermore, Jones (1995) demonstrated that the measure did not reliably assess the negative affect associated with anxiety, suggesting the measure may be assessing only intensity of anxiety but not direction, which could be positive (facilitative anxiety) or negative (debilitative). In response to these weaknesses, a revised version of the CSAI–2 (the CSAI–2R) was published in 2003 (Cox, Martens, & Russell, 2003). This version contains 17 items across the three scales and has held its factor structure through confirmatory factor analysis (Cox et al., 2003). If further validation of this revised instrument continue to support its psychometric properties, the CSAI–2R may prove to represent a better assessment of competitive state anxiety than the CSAI–2 and should be considered.

The SAS

The SAS (Smith, Smoll, & Schutz, 1990) is also among the most widely used assessment instruments in sport psychology and provides another measure of (trait) anxiety possessing strong psychometric properties. Items are presented using a Likert-type scale ranging from 1 (*not at all*) to 4 (*very much*). The SAS has three scales measuring somatic anxiety (nine items), cognitive anxiety–concentration disruption (five items), and cognitive anxiety–worry (seven items). Worry pertains to cognitions regarding consequences of performance, whereas concentration disruption relates to ability to maintain proper focus during competition. Internal consistency reliabilities for the three scales have been satisfactory, ranging across two samples from .74 (cognitive anxiety–concentration disruption) to .92 (somatic anxiety; Smith et al., 1990). Evidence of concurrent validity has been demonstrated through correlations with the SCAT (.47–.80), both scales and total measure as well as evidence for discriminate validity (Smith et al., 1990).

Although there has been documented support for the three-factor structure (Smith et al., 1990), at least two confirmatory factor analyses have demonstrated a superior factor structure when two of the items that were originally intended to load on the cognitive anxiety–concentration disruption scale were moved to the cognitive anxiety–worry scale (Dunn, Causgrove Dunn, Wilson, & Syrotuik, 2000; Prapavessis, Maddison, & Fletcher, 2005). These studies suggest the possible need to adjust the scale items to fit the constructs more appropriately, and caution should be used when interpreting these

factors according to the original scale compositions. Recently, a Sport Anxiety Scale—2 (SAS–2; Smith et al., 2006), has been developed, offering similar scale construction with items not above the fourth-grade reading level.

The TEOSQ

Another construct of pivotal importance in sport and exercise psychology is that of motivation. The TEOSQ (Duda, 1992, 1989; Nicholls, 1989) assesses the tendency to perceive competence in a self-referenced (task) or other-referenced (ego) perspective. These tendencies reflect goal orientations, with individuals possessing high levels of task orientation being more likely to assess their competence through their own mastery of skills, personal development in the sport, and effort provided. Individuals higher in ego goal orientation tend to assess competence through comparison to others, including outscoring or defeating an opponent. Individuals may fall on a continuum from low to high on both qualities; thus, for instance, an athlete may possess both a high task and ego orientation.

The TEOSQ is a 13-item measure assessing task (seven items) and ego (six items) goal orientations in sport. Respondents indicate on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) their amount of agreement to items pertaining to times when they felt successful in sport, such as learning new competencies through effort. The measure has demonstrated adequate internal consistency reliability in a sample of elite skiers, .79 (task) and .81 (ego; Duda & White, 1992). Three-week test–retest reliabilities were .82 (task) and .89 (ego) in a validation sample (Duda, 1989). Scores on the measure also have shown evidence for concurrent and predictive validity across samples (e.g., Duda & White, 1992). Confirmatory factor analysis has supported the two-factor structure (e.g., Guivernau & Duda, 1994).

The GEQ

Another important psychological construct in relation to sport is *group cohesion*. Defined as "a dynamic process that is reflected in the tendency for a group to stick together and remain united in the pursuit of its instrumental objectives and/or for the satisfaction of member affective needs" (Carron et al., 1998, p. 213), cohesion has been assessed by a number of different measures, the most widely used of them being the GEQ (Carron, Widmeyer, & Brawley, 1985).

The instrument assesses cohesion from a multidimensional perspective, measuring the task and social aspects of what attracts individuals to a group as well as their assessment of the group's functioning as a unit (Brawley, Carron, & Widmeyer, 1987). The measure includes four subscales including the individual attractiveness to the tasks of the group, the individual attractiveness to the social aspects of the group, the degree of integration the group has for its tasks, and the degree of integration the group has for social aspects. The measure contains 18 statements to which responses are given on a level of agreement using a Likert-type scale ranging from 1 (*strongly agree*) to 9 (*strongly disagree*). Items include team unity and personal friendships within the group.

Internal consistency findings have varied substantially across samples. For instance, although a study of university students found Cronbach alpha coefficients ranging from .73 to .83 (Carron & Ramsay, 1994), others have found significantly smaller coefficients (Westre & Weiss, 1991). Research also has provided evidence of the predictive, concurrent, and discriminant validity for scores of the measure (Brawley et al., 1987). Findings regarding the factor structure of the GEQ have been mixed (e.g., Li & Harmer, 1996). For instance, although some samples using confirmatory factor analysis have supported the four-factor structure (Li & Harmer), others have not (Sullivan, Short, & Cramer, 2002). As noted by the test's authors, the development of measures is an ongoing process (Brawley et al., 1987). In an effort to improve the internal consistency of the GEQ, the measure was revised to include only the positive wording of items (Eys, Carron, Bray, & Brawley, 2007). However, also as noted by the authors, using the modified version of the GEQ would increase the risk of the influence of response sets, particularly acquiescence.

The ACSI–28

Significant to the assessment of sport psychology is the measurement of psychological skills in sport.

Numerous measures, including many unpublished and proprietary instruments, have been developed and frequently used for these purposes (e.g., Lesyk's Nine Mental Skills). Among the most often used published inventories are the Test of Performance Strategies (TOPS; Thomas, Murphy, & Hardy, 1999) and the ACSI–28 (Smith, Schutz, Smoll, & Ptacek, 1995). The latter is discussed in more detail here.

The ACSI–28 is a revised 28-item version of the original 87-item ACSI. The measure contains seven subscales (four items each) including Coachability, Concentration, Confidence and Achievement Motivation, Coping With Adversity, Freedom From Worry, Goal Setting/Mental Preparation, and Peaking Under Pressure. A Personal Coping Resources score is computed by summing the total of the seven subscales, purporting to provide a global psychological skills indicator. Respondents indicate, on a 4-point Likert-type scale ranging from 0 (*almost never*) to 3 (*almost always*), how often they have particular sport-related experiences such as achievement planning and performance confidence.

Internal consistency reliability of the total measure has been shown to be high (.86) in a sample of athletes, although alpha coefficients for the individual subscales were expectedly smaller with significant variation from .62 (Concentration) to .78 (Peaking) in a sample of athletes (Smith et al., 1995). Test–retest reliabilities have been generally shown to be adequate, with the notable exception of the Coachability subscale (.47), in a sample of male and female athletes. The seven-factor structure of the measure has been replicated by confirmatory factor analysis (Smith et al., 1995). The measure showed mixed evidence of both convergent and discriminant validity in relation to other established measures and constructs across the various scales as well as susceptibility to the social desirability effect (Smith et al., 1995), which warrants consideration when administering the inventory.

Profile of Mood States (POMS) and Test of Attentional and Interpersonal Style (TAIS)

In addition to these sport-specific measures, many other nonsport and exercise specific psychological inventories are also used in both research and practice. In particular, the POMS and the TAIS are two of the most popular inventories in sport psychology, with surveys typically placing these among the most often used inventories by applied sport psychologists (O'Connor, 2004). Although more extensive summaries of these measures can be found elsewhere, a brief review and considerations are presented.

The POMS (McNair, Lorr, & Droppleman, 1971, 1981, 1992) is the most extensively used inventory in sport and exercise psychology measuring present and typical mood states. The measure assesses six mood states including tension-anxiety, depression-dejection, anger-hostility, vigor-activity, fatigue-inertia, and confusion-bewilderment. Although the measure possesses relatively strong psychometric properties (McNair et al., 1992), use in the sport and exercise context has found mixed results. The much-discussed "iceberg profile" (roughly speaking, high amounts of vigor and lower amounts of the other five mood states) has sometimes differentiated athletes from nonathletes but has not reliably distinguished high-performing athletes from low performers (Beedie, Terry, & Lane, 2000). Research suggests that mood may have some relationship with success compared with an athlete's own expected performance, but test users should be knowledgeable about the predictive limitations of this measure.

Another of the most popular measures in sport and exercise psychology is the TAIS (Nideffer, 1976). Not originally developed for use in sport, the measure assesses the attentional style of athletes across two dimensions, width (narrow or broad) and direction (external or internal). Nideffer reports adequate evidence of reliability and validity; however, other investigations of the psychometric properties of the scores of the instrument have revealed questionable evidence of validity (e.g., Ford & Summers, 1992). Nonetheless, the instrument has been reported to be used effectively with athletes (Ostrow, 2000). In addition, particular sport-specific versions of the TAIS also have been developed (e.g., tennis, baseball, softball, riflery), although these measures do not as yet have documentation that report psychometric properties.

This brief review of widely used measures in sport and exercise psychology serves to highlight instruments with reasonably sound construction

and psychometric properties. Although such attention to psychometric validation procedures and the construction of measures with evidence of reliability and validity can be found in numerous existing and newer measures (e.g. Eys, Loughead, Bray, & Carron, 2009), significant issues in assessment of sport and exercise psychology remain.

CURRENT ISSUES

Substantial numbers of measures exist today in the field. Despite the generally improving quality of measures in comparison with that in past decades, a great number of instruments exist without documented evidence of reliability, validity, or norming data. Because of the proliferation of measures, many of which are without evidence of validity, some in the field have expressed that an effective quarantine on new measures may be worth consideration in an effort to devote more attention to the validation of existing measures (e.g., Gill, 2000; O'Connor, 2004; Schutz, 1994). Although there may yet be need for the development of new measures to assess constructs and answer questions not yet addressed by existing instruments, the message from these notable researchers should be taken; using existing measures for study and strengthening the existing measures should be at the forefront of researchers' thought.

A similar issue raised by some in the field (e.g., Gill, 2000) has been the conducting of atheoretical research that reflects the practice, too often relied upon, of beginning with the availability of measures rather than questions worth asking. For instance, the field has seen a continued (albeit reduced) investigation of personality traits through the use of personality inventories with little consideration of contextual and situational factors. This purely trait-based approach, found unfruitful in the 1960s and 1970s, seems to be in part a function of the assessment instrument rather than theory, directing the research. Instead of research driven by popular "known-commodity" assessment instruments, theory should be used to guide these approaches. These investigations should carefully select from the wide range of measures, as well as methods, available that most appropriately address the research question.

Theory also may provide a bridge between research and practice (Gill, 2000). Similar to other areas of applied psychology, sport and exercise psychology evidences a divide between research and practice. Practical considerations often lead consultants and sport psychologists to tools possessing less-than-ideal evidence of validity in their practices. For instance, some consultants may use a subset of items or even a single item from a measure to assess athletes before or during competition for fear they may disrupt athletes, coaches, or teams. This practical, and perhaps necessary at times, approach puts reliability directly at odds with practice (Kimiecik & Blissmer, 1998) and is illustrative of issues related to the larger disconnect between researchers and practitioners seen in both sport and exercise domains.

Another issue in the assessment of sport and exercise psychology is the assumption of linear relationships between variables. One example of this has been the research on the relationship between anxiety and performance. Although theories have, for some time, elucidated nonlinear relationships between performance and anxiety (and arousal), assessment of these relationships has often failed to accurately measure these associations. A summary of research between somatic anxiety and performance, for instance, found inconsistent correlations (Gould et al., 1987). Both design and research tools should account for nonlinear relationships in their use.

The past 2 decades has included greater attention to cross-cultural validation efforts of instruments, including the translation and validation of many instruments into multiple languages (e.g., Tsorbatzoudis, Barkoukis, Sideridis, & Grouios, 2002). Despite such efforts, there remains a major lack of assessments adequately appropriate with respect for client diversity in areas related to gender, race, ethnicity, sexual orientation, age, and physicality (Gill & Kamphoff, 2009). The widespread lack of attention to diversity in the field has been chronicled and demands greater attention in assessment.

Special considerations also should be given to the ethical use of assessment measures (e.g., Vealey & Garner-Holman, 1998). One consideration is the selection of an instrument with adequate psychometric properties. Resources such as Ostrow's *Directory of Tests* (Ostrow, 1996) and Duda's review of

measures (Duda, 1998) have helped bring these qualities to light and have helped test users make more informed decisions.

Aside from selecting a measure with evidence of reliability and validity for its scores, test users should consider several questions when using assessment measures. One consideration is whether the test user possesses the adequate training and competence, including cultural competence, to properly administer, score, and interpret the measure. The measures themselves vary as to the amount of training necessary to adequately perform such tasks (see McCann et al., 2002), and users should have sufficient knowledge of the measure to prevent common issues such as the inappropriate use of measures, misinterpretation of results, and failure to consider cultural considerations in test use. For instance, self-report measures as administered in sport and exercise psychology, are prone to influence through response sets such as acquiescence and social desirability (e.g., Carron et al., 1985; Cox, 2006), and test users should be knowledgeable about proper administration and interpretation of these inventories. Furthermore, the extent and limits of confidentiality should be addressed with clients before test administration, and timely and effective feedback should be provided after scoring.

In an attempt to address and regularize these ethical concerns, the Association for Applied Sport Psychology published principles and standards regarding the ethical use of assessment, including practice and the conduction of research. These principles and standards serve as an instructive resource for researchers and practitioners alike. However, these guidelines have not become uniform throughout sport and exercise psychology, and further dissemination, training, and application of these ethical principles and standards remains a need in the field.

These significant issues aside, assessment in sport and exercise psychology has made major advancements in the past 2 decades. Continued progress in the field has resulted in greater organization and more informed selection of a group of measures of improving psychometric quality as well as more appropriate use in practice and research. The domain of exercise psychology has seen particular development in the recent years and continues to receive greater attention in research and applied settings. The recent progress and development of the field provides reason for optimism for the continued development and enhancement of assessment in sport and exercise psychology.

References

American Psychological Association. (2010). *Becoming a sport psychologist*. Washington, DC: Author. Retrieved from http://www.apa.org/divisions/div47/APA%20Div%2047%20(1)/about/about_becoming-sportpsych.html

Beedie, C., Terry, P., & Lane, A. (2000). The profile of mood states and athletic performance: Two meta-analyses. *Journal of Applied Sport Psychology*, *12,* 49–68. doi:10.1080/10413200008404213

Brawley, L., Carron, A., & Widmeyer, W. (1987). Assessing the cohesion of teams: Validity of the Group Environment Questionnaire. *Journal of Sport Psychology*, *9,* 275–294.

Carron, A., Brawley, L. R., & Widmeyer, W. N. (1998). The measurement of cohesiveness in sport groups. In J. L. Duda (Ed.), *Advances in sport and exercise psychology measurement* (pp. 213–226). Morgantown, WV: Fitness Information Technology.

Carron, A., & Ramsay, M. (1994). Internal consistency of the Group Environment Questionnaire modified for university residence settings. *Perceptual and Motor Skills*, *79*, 141–142. doi:10.2466/pms.1994.79.1.141

Carron, A., Widmeyer, W., & Brawley, L. (1985). The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. *Journal of Sport Psychology*, *7*, 244–266.

Cox, R. H. (2006). *Sport psychology: Concepts and applications* (6th ed.). Boston, MA: McGraw-Hill.

Cox, R. H., Martens, M. P., & Russell, W. D. (2003). Measuring anxiety in athletics: The revised Competitive State Anxiety Inventory-2. *Journal of Sport and Exercise Psychology, 25*, 519–533.

Cozby, P. C. (2007). *Methods in behavioral research* (9th ed.). Boston, MA: McGraw-Hill.

Duda, J., & White, S. (1992). Goal orientations and beliefs about the causes of sport success among elite skiers. *The Sport Psychologist*, *6*, 334–343.

Duda, J. L. (1989). The relationship between task and ego orientation and the perceived purpose of sport among male and female high school athletes. *Journal of Sport and Exercise Psychology, 11*, 318–335.

Duda, J. L. (1992). Motivation in sport settings: A goal perspective approach. In G. Roberts (Ed.), *Motivation in sport and exercise* (pp. 57–91). Champaign, IL: Human Kinetics.

Duda, J. L. (1998). *Advances in sport and exercise psychology measurement*. Morgantown, WV: Fitness Information Technology.

Dunn, J. G. H., Causgrove Dunn, J., Wilson, P., & Syrotuik, D. G. (2000). Reexamining the factorial composition and factorial structure of the Sport Anxiety Scale. *Journal of Sport and Exercise Psychology, 22*, 183–193.

Eys, M., Carron, A., Bray, S., & Brawley, L. (2007). Item wording and internal consistency of a measure of cohesion: The Group Environment Questionnaire. *Journal of Sport and Exercise Psychology*, *29*, 395–402.

Eys, M., Loughead, T., Bray, S., & Carron, A. (2009). Development of a cohesion questionnaire for youth: The Youth Sport Environment Questionnaire. *Journal of Sport and Exercise Psychology*, *31*, 390–408.

Fogarty, G. (1995). Some comments on the use of psychological tests in sport settings. *International Journal of Sport Psychology*, *26,* 161–170.

Ford, S., & Summers, J. (1992). The factorial validity of the TAIS attentional-style subscales. *Journal of Sport and Exercise Psychology*, *14,* 283–297.

Fung, L., Ng, J., & Cheung, S. (2001). Confirmatory factor analysis of the Trait Sport-Confidence Inventory and State Sport-Confidence Inventory on a Chinese sample. *International Journal of Sport Psychology*, *32*, 304–313.

Gill, D. (2000). *Psychological dynamics of sport and exercise* (2nd ed.). Champaign, IL: Human Kinetics.

Gill, D., & Kamphoff, C. (2009). Cultural diversity in applied sport psychology. In R. J. Schinke & S. J. Hanrahan (Eds.), *Cultural sport psychology* (pp. 45–56). Champaign, IL: Human Kinetics.

Gould, D., Petlichkoff, L., Simons, J., & Vevera, M. (1987). Relationship between Competitive State Anxiety Inventory—2 subscale scores and pistol shooting performance. *Journal of Sport Psychology*, *9,* 33–42.

Guivernau, M., & Duda, J. L. (1994). Psychometric properties of a Spanish version of The Task and Ego Orientation in Sport Questionnaire (TEOSQ) and Beliefs about the Causes of Success Inventory. *Revista de Psicología del Deporte, 5*, 31–51.

Isogai, H., Brewer, B., Cornelius, A., Komiya, S., Tokunaga, M., & Tokushima, S. (2001). Cross-cultural validation of the Social Physique Anxiety Scale. *International Journal of Sport Psychology*, *32,* 76–87.

Jones, G. (1995). More than just a game: Research developments and issues in competitive anxiety in sport. *British Journal of Psychology, 86*, 449–478. doi:10.1111/j.2044-8295.1995.tb02565.x

Kimiecik, J. C., & Blissmer, B. (1998). Applied exercise psychology: Measurement issues. In J. L. Duda (Ed.), *Advances in sport and exercise psychology measurement* (pp. 447–460). Morgantown, WV: Fitness Information Technology.

Landers, D. (1995). Sport psychology: The formative years, 1950–1980. *The Sport Psychologist*, *9*, 406–417.

Lane, A. M., Sewell, D. F., Terry, P. C., Bartram, D., & Nesti, M. S. (1999). Confirmatory factor analysis of the Competitive State Anxiety Inventory—2. *Journal of Sports Sciences, 17*, 505–512. doi:10.1080/026404199365812

Lavallee, D., Kremer, J., Moran, A. P., & Williams, M. (2004). *Sport psychology: Contemporary Themes*. London, England: Palgrove MacMillan.

Li, F., & Harmer, P. (1996). Confirmatory factor analysis of the Group Environment Questionnaire with an intercollegiate sample. *Journal of Sport and Exercise Psychology*, *18,* 49–63.

Mahoney, M. J. (1989). Sport psychology. In I. S. Cohen (Ed.), *The G. Stanley Hall lecture series* (Vol. 9, pp. 101–134). Washington, DC: American Psychological Association.

Marsh, H. W. (1998). Foreward. In J. L. Duda (Ed.), *Advances in sport and exercise psychology measurement* (pp. xv–xix). Morgantown, WV: Fitness Information Technology.

Martens, R. (1979). About smocks and jocks. *Journal of Sport Psychology*, *1,* 94–99.

Martens, R., Vealey, R. S., Burton. D., Bump, L., & Smith, D. E. (1990). Development and validation of the Competitive State Anxiety Inventory—2. In R. Martens, R. S. Vealey, & D. Burton (Eds.), *Competitive anxiety in sport* (pp. 117–178). Champaign, IL: Human Kinetics.

McCann, S. C., Jowdy, D. P., & Van Raalte, J. L. (2002). Assessment in sport and exercise psychology. In J. L. Van Raalte & B.W. Brewer (Eds.), *Exploring sport and exercise psychology* (2nd ed., pp. 291–305). Washington, DC: American Psychological Association.

McNair, P. M., Lorr, M., & Droppleman, L. F. (1971). *Profile of Mood States manual*. San Diego, CA: Educational and Industrial Testing Service.

McNair, P. M., Lorr, M., & Droppleman, L. F. (1981). *Profile of Mood States manual* (2nd ed.). San Diego, CA: Educational and Industrial Testing Service.

McNair, P. M., Lorr, M., & Droppleman, L. F. (1992). *Revised manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

Mischel, W. (1968). *Personality and assessment*. Hoboken, NJ: Wiley.

Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.

Nideffer, R. (1976). Test of attentional and interpersonal style. *Journal of Personality and Social Psychology*, *34*, 394–404. doi:10.1037/0022-3514.34.3.394

O'Connor, E. (2004). Which questionnaire? Assessment practices of sport psychology consultants. *The Sport Psychologist*, *18,* 646–648.

Ostrow, A. C. (1996). *Directory of psychological tests in the sport and exercise sciences.* Morgantown, WV: Fitness Information Technology.

Ostrow, A. C. (2000). Sport psychology: Assessment. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (7th ed., pp. 449–452). Washington, DC: American Psychological Association.

Prapavessis, H., Maddison, R., & Fletcher, R. (2005). Further examination of the Factor Integrity of the Sport Anxiety Scale. *Journal of Sport and Exercise Psychology*, *27,* 253–260.

Schutz, R. W. (1994). Methodological issues and measurement problems in sport psychology. In S. Serpa, J. Alves, & V. Pataco (Eds.), *International perspectives on sport and exercise psychology* (pp. 35–57). Morgantown, WV: Fitness Information Technology.

Smith, R., Schutz, R., Smoll, F., & Ptacek, J. (1995). Development and validation of a multidimensional measure of sport-specific psychological skills: The Athletic Coping Skills Inventory-28. *Journal of Sport and Exercise Psychology*, *17,* 379–398.

Smith, R., Smoll, F., Cumming, S., & Grossbard, J. (2006). Measurement of Multidimensional Sport Performance Anxiety in Children and Adults: The Sport Anxiety Scale—2. *Journal of Sport and Exercise Psychology*, *28,* 479–501.

Smith, R., Smoll, F., & Schutz, R. (1990). Measurement and correlates of sport-specific cognitive and somatic trait anxiety: The Sport Anxiety Scale. *Anxiety Research*, *2,* 263–280. doi:10.1080/08917779008248733

Sullivan, P. J., Short, S. E., & Cramer, K. M. (2002). Confirmatory factor analysis of the Group Environment Questionnaire with co-acting sports. *Perceptual and Motor Skills*, *94,* 341–347. doi:10.2466/pms.2002.94.1.341

Thomas, P. R., Murphy, S. M., & Hardy, L. (1999). Test of Performance Strategies: Development and preliminary validation of a comprehensive measure of athletes' psychological skills. *Journal of Sports Sciences, 17*, 697–711. doi:10.1080/026404199365560

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology*, *9,* 507–533. doi:10.2307/1412188

Tsorbatzoudis, H., Barkoukis, V., Sideridis, G., & Grouios, G. (2002). Confirmatory factor analysis of the Greek version of the Competitive State Anxiety Inventory—2 (CSAI–2). *International Journal of Sport Psychology*, *33,* 182–194.

Vealey, R. S., & Garner-Holman, M. (1998). Applied sport psychology: Measurement issues. In J. Duda (Ed.), *Advances in sport and exercise psychology measurement* (pp. 433–466). Morgantown, WV: Fitness Information Technology.

Westre, K. R., & Weiss, M. R. (1991). The relationship between perceived coaching behaviors and group cohesion in high school football teams. *The Sport Psychologist, 5*, 41–54.

Williams, J. M., & Straub, W. F. (2009). Sport psychology: Past, present, future. In J. M. Williams (Ed.), *Applied sport psychology* (pp. 1–17). Boston, MA: McGraw-Hill.

CHAPTER 32

PSYCHOLOGICAL ASSESSMENT WITH OLDER ADULTS

Tammi Vacha-Haase

Everyone ages, albeit in an individual and unique way, with varying degrees of cognitive decline, emotional distress, and decreased physical functioning. Similar to other developmental stages across the life span, those over the age of 65 can be described on a continuum—from successful, to normal, to pathological. The later years of life for some older adults appear easy as they experience good health, maintain their cognitive functioning, and cope with emotional distress. At the other end of the spectrum, another subgroup will experience "pathological" levels of cognitive deficits, emotional turmoil, and poor physical health that will negatively affect their quality of life. Although many older adults will experience obstacles, the majority do not develop psychological disorders (Scott et al., 2008) and fall within the "normal" range of the aging process (Ayis, Paul, & Ebrahim, 2010).

Whether for research purposes or for diagnostic and treatment planning, the assessment goal is to increase understanding of the older adult's affect, behavior, personality traits, general functioning, and overall well-being. Unfortunately, this can prove to be quite difficult. Perhaps the challenge of psychological assessment with older adults can be attributed to the heterogeneity of the population as well as the rates of comorbidity of cognitive, physical, emotional, and functional aspects of this population. Although not synonymous with age, older adults are at increased risk for medical conditions (Centers for Disease Control and Prevention, 2007). This may directly affect the testing session, as well as present a more complicated medical history, that may (or may not) be a contributing (or unknown) factor. Differential diagnosis becomes increasingly important regarding questions of delirium or dementia as well as the degree of physical, psychological, chronic, situational, or environmental symptoms.

Another obstacle may be the relatively low numbers of assessment instruments that include norms for those over 65 years of age, and even fewer that are developed specifically for older adults. Years ago, Butler (1975) described the area of aging as "the neglected stepchild of the human life cycle" (p. 126). Today, with the relatively limited assessment measures available for those in their later years, perhaps geriatric assessment can best described as a recently added stepchild of the assessment area. The future holds great promise, however, as a quick review of the PsycINFO database suggests that when compared with earlier years, the previous decade brought about almost twice the number of publications in the area of psychological assessment with an older adult population.

ASSESSMENT APPROACH WITH OLDER ADULTS

Psychological assessment with older adults builds on the principles and techniques required in general assessment. The referral questions must be clarified, followed by the selection of tests, with thought given to psychometric adequacy, relevance, incremental validity, and complementary functions. Issues of integration of data sources, interpretation of results, and feedback approach arise, as do ethical

DOI: 10.1037/14048-032
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

and legal requirements. Thus, expertise in the basics of assessment must be relied on but with an additional skill set incorporated, including knowledge of the aging process and adjusted clinical skills (Knight, 2004).

Clinical Adjustments

In general, there should be appropriate adjustments for the older adult population with the purpose of modifying test procedures to increase the collection of worthwhile data, especially when physical disabilities or motor difficulties are present (Schlenoff, 1989). Although standardization of testing procedures has long been the norm within the assessment process, modifications may be required to fit the unique situation of the older adult. A thoughtful, tailored approach to test administration may be called for. However, caution should always be at the forefront, and any modifications or accommodations should be made with considerable thought and analyses.

Other needs for clinical adjustment may be apparent before beginning the assessment, such as flexibility in making the appointment arrangements, time of day, availability of transportation, or accessibility of location. Because chronic illness often accompanies older age, as well as lowered stamina, fatigue, or pain issues, testing sessions may need to be scheduled for shorter time periods, across a longer period of time, or allow for increased breaks. During the testing periods, extra care should be given to the environment to encourage optimal performance, including a space that is well lit, adequately ventilated, soundproof, and a comfortable temperature, as older adults may be more sensitive to negative conditions. In addition, diminished sensation may become an issue, as one in three people older than age 60 and half of those older than 85 have hearing loss, and more than half suffer from low vision. In these situations, modifications such as larger text or use of a lower tone and slower rate of speech may be needed.

The older adult's overall comfort level with the assessment process may also require an adjustment in establishing rapport, with an increased and possibly extended focus on developing a supportive and working relationship. Some older adults will be unfamiliar with a testing situation, either having limited experience with any type of testing or having lived many decades since they were in a school setting. They may find the testing process unknown and disconcerting, if not outright distressful, causing them to present as hesitant, cautious, or even suspicious. Depending on the circumstances, some older adults may be frightened, embarrassed, or anxious regarding their participation in psychological assessment.

Providing Feedback

Although providing feedback is a critical aspect of psychological assessment, little research exists regarding preference of feedback (Brenner, 2003). More recently, research has supported that psychological assessment procedures, when combined with personalized, collaborative, and highly involving test feedback, have positive, clinically meaningful effects (Poston & Hanson, 2010). However, with older adults, feedback can take on an increased challenge, requiring expertise of combining effective clinical skills when working with older adults and the understanding of psychological assessment and outcomes.

The success of a feedback session includes maintaining rapport and being cognizant of the older adult's understanding of the information being provided. General goals would include explaining strengths and weakness, application to daily life, and changes that might be expected in the future; inclusion of potential issues of safety is fundamental. Feedback may explore possible treatments, required additional assessments, or other needed follow-up. Focus should be given to physical health as well as psychosocial and environmental aspects, including how this might affect the older adult and significant others, such as a spouse, family member, caregiver, or other health care provider.

A tailored approach to feedback is most likely the best, as interest in or desire for feedback may be different for each older adult. For example, one older adult may decline to schedule a feedback appointment because of transportation difficulties, additional cost, or feeling overwhelmed at attending another appointment. Another older adult may have little interest in the results and request that all

information be addressed to a family member or other health care provider. Thus, before conducting an assessment, feedback options should be clarified, including (a) who should receive the results, (b) the format, (c) the timing, and (d) the method of delivery.

Cultural Aspects

Regardless of the age group, cultural attitudes, beliefs, and practices inform all aspects of the psychological assessment process. However, placing older adults in their cultural context provides information on how they live as well as their view of health, illness, and healing; end of life, and death; and the role of the family and health care providers. The diversity of the aging population offers a unique challenge to understand better the role of race, culture, and ethnicity in psychological assessment. Direct issues such as literacy, language, and education level are at play, but so are subtle issues, including level of acculturation, racial socialization, and experiences. Awareness must be expanded to understand

> how individual diversity in all of its manifestations (including gender, age, cohort, ethnicity, language, religion, socioeconomic status, sexual orientation, gender identity, disability status, and urban/rural residence) interacts with attitudes and beliefs about aging, to utilize this awareness to inform their assessment and treatment of older adults. (Knight et al., 2009, p. 208)

Level of comfort, confidence in ability, or approach to answering questions during the assessment session may vary among older adults from differing groups. For example, an older man with a traditional view of masculinity may approach the assessment process utilizing well-learned masculine scripts (Mahalik, Good, & Englar-Carlson, 2003), being a silent or "strong and tough" male. Rather than assessing this type of behavior as negative, in this example the older man might best be understood through a combination of gender, age, and individual personality characteristics (Vacha-Haase, Wester, & Christianson, 2010). In addition, attention must be directed to the integration of multiple identities (Yakushko, Davidson, & Williams, 2009); for example, is this older man's gender his primary identity and the starting point for understanding his worldview? Or is it his age? His ethnicity? Success may be best met through recognition of the multitude of his roles or identities; that is, understanding this individual as a 78-year-old African American man who is originally from the South; is heterosexual, Christian, widowed; has a high school diploma and 20 years in the military; is able bodied but with diabetes and has been prescribed medication for high blood pressure; is employed part time; and values his role as a father and grandfather. Manly (2006) cautioned against assumptions and recommended asking older adults directly to identify their ethnicity, where they grew up and their cultural experiences, and details about their education. Physical changes and psychological functioning placed within the context of gender role socialization and aging are fundamental aspects to the assessment process.

The reader is encouraged to become familiar with the literature regarding ethnic differences in older adult populations (e.g., Baden & Wong, 2008; Baird, Ford, & Podell, 2007) as well as for specific ethnic minority or other groups (e.g., Native American [Ferraro, 2001]; Latino/Hispanic [Nuevo, Mackintosh, Gatz, Montorio, & Wetherell, 2007], men [Vacha-Haase, Wester, & Christianson, 2010], and sexual orientation [Brick, Lunquist, Sandak, & Taverner, 2009]). An excellent overview infusing the literature in multicultural and geropsychology is provided in *Multicultural Competency in Geropsychology* (APA Committee on Aging, 2009).

Cohort Effect

The assessment area has been critiqued as having a European American orientation; this may also be extended to having a youth basis. Thus, embedded within cultural consideration when working with older adults also lies the aspect of their cohort group.

The relationship between the year of birth and historical period has long been noted and is often referred to as a birth cohort or generational cohort. Cohort can be thought of as the aggregate of persons born in the same time interval (Ryder, 1965), with the rationale that attitudes are shaped in youth and

remain stable after a certain age. Thus, a cohort is a birth-year-defined group that is "socialized into certain abilities, beliefs, attitudes" distinguishing them from other age groups (Laidlaw & Knight, 2008, p. 61). Because historical forces and social norms of the times are influential, the cohort group provides a framework for variations in education, nuances in word meanings and usages, values, social norms, and role expectations, all of which have the potential to influence the assessment process and potential findings.

Interdisciplinary Approach

Interaction with other disciplines is prominent in psychological assessment with older adults because consultation skills and the ability to succinctly describe results to professionals from various disciplines become paramount. Ensuring the usefulness of psychological assessment for other professionals (e.g., physician, nurse, social worker, occupational therapist) involved in the older adult's care cannot be overemphasized (Brenner, 2003).

Depending on the setting, there may be a demand for shorter turn around time, with the assessment being more targeted and problem-solving oriented. At times, the assessment may need to fit within the framework of the medical model, with focus given to aspects most relevant to treatment planning and assessment of risk management. The current professional climate also focuses on time and cost efficiency, with an increase in accountability and decisions made less on clinical judgment and more on empirically derived decisions (Hunsley & Mash, 2008).

Edelstein, Martin, and Koven (2003) identified two principal assessment paradigms with older adults: traditional and behavioral. Although there is overlap, in general, the traditional paradigm focuses on trait-oriented personality characteristics, intelligence, and diagnosis; dispositional characteristics are inferred from self-reports of feelings, attitudes, and behaviors as well as observations. The question of interest is, "Why does the person act this way?" In providing an answer, the older adult may be asked to complete a depression inventory, a measure of anxiety, and a brief cognitive screening tool.

In contrast, a behavioral approach explores the social learning components by explaining the person's behavior through the conditions or circumstances in which the behaviors occur; this is known as *environmental determinism*, or the premise that behavior is functional and thus emitted in response to an environmental cue. Unlike the traditional approach where comparison of the older adult's level of anxiety with population norms would be assessed, the behavioral approach would see this as a single case and seek to assess the target behavior for that individual through direct observation in various situations and environments. Rather than asking "why?" instead the question focuses on, "When and where does the person behave this way?" with focus given to the circumstances when the older adult is exhibiting the target behavior. Options for assessment might include behavioral observations by a trained technician who records target behaviors and possible causal factors; self-monitoring with systematic recording of one's own behaviors; questionnaires; and experimental manipulations, either in natural settings or specifically identified environments.

Although there are notable differences between these two paradigms, they can be combined to provide an effective psychological assessment. Multimethod as well as multidimensional assessment may be optimal in the majority of testing situations with older adults (Groher, 1989; Schlenoff, 1989).

Settings

Psychological assessment with older adults is completed for several reasons in a variety of settings. Although many locations are for all ages (e.g., medical hospital, psychiatric hospital, and independent practice office), there are also settings that may be somewhat unique to an older population, including assisted living, long-term care (LTC) facilities, and hospice. Regardless of the referral question, the setting will be an important factor in the overall assessment. For example, the psychological assessment of a 77-year-old man living independently in his own home, a 77-year-old man living in a LTC facility, and a 77-year-old man in hospice would each require a different testing approach, even with a similarly identified referral question.

The majority of older adults live independently in the community, and would most likely complete a psychological assessment, if needed, through a

community mental health center, independent practice, or medical practice. However, more than 40% of those over age 65 will enter a LTC facility sometime during their life (Kemper & Murtaugh, 1991), which may be where the psychological assessment is completed. For a thorough review of assessment in LTC facilities, readers are referred to Edelstein, Northrop, and MacDonald (2009).

Ethics

Ethical practice of psychological assessment with an older adult population starts with the underlying ethical standards of the profession (APA, 2010). Certainly the four core factors for all ethics are applicable: respect for autonomy, nonmalefience, beneficence, and justice. However, given the specifics of the population, complex dilemmas can arise. Here again, competence and general assessment skills are required but not sufficient.

When working with older adults, identifying or consistently clarifying who is the client frequently surfaces, as others often become involved in the care. Whether it be a concerned family member or a care provider or another health care provider, within any assessment, clarifying "who is the client" is significant. Related are the limits of confidentiality in the assessment process. Depending on the situation, confidentiality may or may not be guaranteed if the results of the psychological assessment are directly added to the medical chart. Although each circumstance will differ depending on the purpose of the evaluation, setting, and overall situation, emphasis on "do no harm" and "respect for autonomy" is recommended.

Another issue in working with older adults comes within the area of informed consent, in both treatment and research (Dunn & Misra, 2009). Because of cognitive limitations, hesitancy or fear, or delirium, the older adult may not be able to understand the nature of the assessment, or the potential consequences of not participating in the assessment process (Etchells et al., 1999).

COMMON AREAS OF PSYCHOLOGICAL ASSESSMENT

Although there is perhaps no more heterogeneous population than that of those over the age of 65, a relatively small number of general areas form the basis of need for psychological assessment in an older adult population. Some, such as mood and personality, are similar to other ages, but the uniqueness comes in the approach to the issue. Other areas, such as capacity and daily functioning are more common to those as they age.

Developmental or age-specific issues will often be relevant in the assessment process. That is, psychosocial factors that are most prominent during the later years of life may often emerge, including family dynamics, available support systems, recent transitions such as retirement, or death and loss, such as the death of a spouse or family member. Assessment may also include areas such as adjustment to new roles, and social expectations as well as cognitive and physical changes.

Functional Assessment

Assessment of cognitive functioning with older adults ranges from a brief screening to more in-depth exploration of memory and executive functioning. Although cognitive functioning fits more closely within the premises of neuropsychological assessment (see Bush & Martin, 2005, for a review), functional assessment is a related concept to cognitive ability that may fall under psychological assessment. In their comprehensive review on the measurement of functional capacity, Patterson and Mausbach (2010) defined functional capacity as "an individual's capability, under controlled conditions, to perform tasks and activities that are necessary or desirable in his or her life" (p. 140). Although influenced by both cognitive and physical ability, functional capacity represents basic living skills and abilities that enable a person to go about their day, often within the context of the older person's ability to live independently. This tends to be divided into two broad areas, including basic self-care behaviors (e.g., personal hygiene) and skills requiring cognitive complexity such as planning or involving completion of multiple steps (e.g., managing finances, shopping, doing laundry). In a chapter on best practices of functional assessment with older adults, Kresevic (2008) identified the value of current assessment, highlighting the potential for being able to provide preventative treatment as well as identifying decline.

Given the gravity of decisions often made on these assessment outcomes, measuring functional capacity has received increased attention and is viewed as a growing area of need (Patterson & Mausbach, 2010). Previously an area that was somewhat neglected, more than 50 different assessment instruments have emerged. Instruments tend to be self-report of the older adult and/or the caregiver, performance based, and clinician ratings.

Capacity

Another emerging area of practice (Moye & Marson, 2007) with fairly high-stakes outcome assessment is that of capacity (Attix & Welsh-Bohmer, 2006). A relatively complex construct, capacity may be best described as including the understanding, appreciation, reasoning and expressing a choice (Grisso & Appelbaum, 1998). Although focus is often on the cognitive effect of decision making, there are also medical factors that can affect capacity. These include infections, endocrine disorders, cardiovascular disease, chronic obstructive pulmonary disease, renal disease, dehydration and malnutrition, and chronic pain.

Currently, there is no gold standard for acceptable criteria or an agreed-upon approach to operationalizing capacity (Edelstein, 2000). In addition, capacity is situational and contextual and thus assessed in the context of specific functions at a given time. Because decision making tends to reflect Western culture of self-determinism, overall assessment should be viewed within the context of cultural influences related to ethnicity, gender, region, and with respect of the values and interests of the older adult (Moberg & Rick, 2008).

Although one approach may be to assess an older adult's overall ability for everyday decision making (Lai & Karlawish, 2007), in general, there are four areas of capacity: (a) *medical*, the ability to consent to medical treatment; (b) *financial*, the ability to manage financial affairs, including a broad range of complex set of abilities as well as specific skills required for money management and making financial decisions; (c) *contractual*, the ability to enter into a contract; and (d) *testamentary*, the ability to make a will. For further reading about capacity and assessment, the book edited by Qualls and Smyer (2007) is recommended.

Mood

Psychological assessment of older adults with respect to mood may include level of current adjustment, depression (see a review by Fiske, Wetherell, & Gatz, 2009), and anxiety (see a review by Stanley & Beck, 2000). Although affect is the focus, a cognitive component remains present with older adults (Steffens & Potter, 2008) because impaired thinking is often present in depression and anxiety among older adults, and the combination raises the risk for a number of adverse emotional and biological outcomes. The high extent to which older adults experience a combination of anxiety and depression with physical medical conditions causes a complex diagnostic puzzle. In fact, Kim, Braun, and Kunik (2001) stated that an "astute clinician should always consider medical causes in any presentation of acute or chronic depression or anxiety" (p. 121), ruling out medical causes such as cardiovascular or pulmonary disease. In addition, pain, a common factor in older age, may be present and having a direct effect on mood. An excellent review of pain assessment is offered by Turk, Okifuji, and Skinner (2008).

Adding to the complexity of mood assessment is the need to differentiation between depression and grief or bereavement as well as the risk of suicide. Unfortunately, the majority of assessment tools were not developed or standardized with older adult samples, and most measures of late-life depression and hopelessness do not include items assessing suicidal features (Heisel & Duberstein, 2005). An excellent review of suicide risk assessment is offered in Duberstein and Heisel (2008); and for those conducing assessments with older adults in LTC, Reiss and Tishler's work (2008a, 2008b) is recommended.

The aforementioned work may help to explain the differences within the literature regarding prevalence of mood disorder in older adults. For example, although older women are twice as likely to experience depression (Heo, Murphy, Fontaine, Bruce, & Alexopoulos, 2008), a recent study of 1,900 participants revealed that depressive symptoms appear to be experienced at fairly uniform levels across the age span by both genders and among Whites, Blacks,

and Hispanics (Bracken & Reintjes, 2010). Random-sample epidemiological surveys have reported a wide range of diagnosable anxiety among older adults living in the community, ranging from 1% to 9% and possibly as high as 12% in primary care.

A number of assessment instruments are available to assess depression with older adults, with some being more or less suitable based on setting, cognitive ability, and time allotted. An excellent review of psychological assessment of depression in later life is provided by Fiske and O'Riley (2008). There are also several assessment instruments to measure anxiety; however, in a review of assessment of anxiety in late life, Kogan, Edelstein, and McKee (2000) warned that the majority of available assessment instruments "lack sufficient evidence for their psychometric soundness for use with older adults" explaining "most of the studies examining the psychometric properties of these instruments have methodological shortcomings (i.e., homogeneity of samples, small sample sizes, restriction of age ranges, concerns regarding method variance) that limit the usefulness of their findings" (p. 127). Although recent research (e.g., Segal et al., 2010) highlights promising new measures in the area of age-specific assessment measures for anxiety, currently there is no single measure of anxiety that performs adequately in screening, measuring severity, and monitoring changes (Dennis, Boddington, & Funnell, 2007).

Personality

Longitudinal studies suggest an overall picture that personality traits are substantial stable across the life span. However, more recent research (Schaie, 2005) suggested that with, age, both neuroticism and agreeableness tend to increase, with a decrease in extraversion and conscientiousness. Of course, non-normative changes in personality can occur, as life experiences affect emotional well-being. Perhaps the death of an adult child, the need to assume a caregiver role for a spouse, or other significant life experiences can change an individual's personality. A number of medical events, such as a stroke, can also cause someone's personality to take on a different presentation style.

Similar to other ages, personality assessment is complex and requires adequate information from multiple sources, including personality assessment instruments. For example, older adults are not immune from the concept of social desirability and the role it has played in the history of personality assessment (Helmes, 2000).

Currently, a personality assessment measure developed specifically for those in later life does not exist, even though it has been argued that there are unique personality aspects within the older adult population. This, unfortunately, also limits the use of available assessment instruments as well as current diagnostic criteria for older adults with personality disorders (Balsis, Segal, & Donahue, 2009). Readers are directed to the book edited by Rosowsky, Abrams, and Zweig (1999) for further reading.

Substance Abuse

Older adult substance use and addiction to alcohol and drugs (legal and illegal) are a growing concern. Although alcohol and substance abuse is the third leading health problem among Americans 55 years of age and older, only in the past decade has the recognition of the pervasiveness of substance abuse problems among older adults come to light (Fingerhood, 2000). The lack of awareness, possibly because of ageism of professionals as well as denial among baby boomers, often serves as a barrier (Stewart & Oslin, 2001) for older adults to receive treatment as well as appropriate assessment.

According to the annual National Survey on Drug Use and Health sponsored by the Substance Abuse and Mental Health Services Administration (2009), half of those between the ages of 60 to 64 used alcohol in the past month, and almost 40% of individuals 65 and older used alcohol in the past month. Six percent of adults aged 65 and older reported binge drinking, and one out of 50 reported heavy drinking. In a study of Medicare recipients, Merrick et al. (2008) found that 9% of community-dwelling older adults reported unhealthy drinking. Assessment of substance use may be most effective provided through general medical settings such as primary care (Stewart & Oslin, 2001); unfortunately, family practice physicians do not routinely screen their older adults patients for alcohol use (Sharp & Vacha-Haase, 2010).

Quality of Life and Others

Although there is no one widely accepted theoretical framework for quality of life or a general consensus concerning which areas are necessary for a comprehensive definition (Halvorsrud & Kalfoss, 2007), there is some agreement that quality of life should be understood from the older adult's individual perspective as well as multidimensional in nature, including physical, emotional, and social domains. Research regarding quality of life greatly expanded in the 1990s, resulting in more than 100 definitions (Cummins, 1997) and possibly more than 1,000 measures of various aspects of quality of life (Hughes & Hwang 1996). An excellent review of empirical studies investigating the conceptualization and measurement of quality of life is offered by Halvorsrud and Kalfoss (2007).

In addition to quality of life, there are several other areas that would seem relevant for psychological assessment with older adults. These include topics such as behavioral actions (e.g., wandering, aggression), sleep disorders, driving ability, chronic pain, and fall predictors. However, because of space limitations, readers are referred to comprehensive resources covering assessment and treatment for emotional disorders (Laidlaw & Knight, 2008), psychopathology (Whitbourne, 2000), physical and mental health (Cavanaugh, Cavanaugh, Qualls, & McGuire, 2010), rehabilitation (Lichtenberg & Schneider, 2010), and neuropsychology (Attix & Welsh-Bohmer, 2006).

Standards for Geriatric Psychological Assessment

There is a growing body of literature designed to assist in appropriate psychological assessment with older adults. This includes the original guidelines for practitioners working with older adults (APA Working Group on the Older Adult, 1998) as well as more recent recommendations (Molinari et al., 2003), aspirational competencies for professional geropsychology (Knight et al., 2009), and, the infusion of multicultural competencies (APA Committee on Aging, 2009). For assessment with older adults experiencing cognitive decline, the handbook *Assessment of Older Adults With Diminished Capacity: A Handbook for Psychologists* (American Bar Association & APA, 2008) and recently revised guidelines for evaluation of older adults with cognitive impairment (APA Task Force to Update the Guidelines for the Evaluation of Dementia and Age-Related Cognitive Decline, 2012) are highly recommended.

Although knowledge is emerging regarding the selection of instruments, alterations in test administration, and interpretation of test data, the reality remains that there are no easy answers. Agreement has not been reached, nor may it ever be possible, to identify which one assessment instrument best answers a specific referral question or is most appropriate with certain older adults or in a given geriatric setting. There has been progress, however, because the command to "gather the maximum amount of information in a short period of time" (Groher, 1989, p. 109) does not reflect the full extent of the consideration given to psychological assessment with older adults. Edelstein and his colleagues (Edelstein et al., 2008) as one example, provided an overview of the most commonly used assessment instruments in five clinical domains, including depression, anxiety, suicide ideation, sleep disorders, and personality. Exhibit 32.1 provides as extensive list as possible of assessment instruments in the areas of cognition, mood, and personality.

In the selection of a test for use with older adults, psychometric properties of each assessment instrument must be considered. Of course, no test score has perfect reliability or validity. However, psychometric properties, including a relatively newer concept of psychometric fairness (Geisinger, Boodoo, & Noble, 2002) must always be included in the selection process.

Just as with any other group, high-quality norms are vital to high-quality clinical practice and psychological assessment. If norms are not comparable, results can lead to misdiagnosis, with the potential to be deficit oriented, overidentifying pathology and, thus, reducing the clinical meaningfulness of the assessment. Unfortunately, many of the tests used in the assessment of older adults were not developed specifically for an older population, nor do they have age-appropriate norms. Although there has been improvement in the past years, relatively few assessments have been normed on those over age 65, and even fewer for those in their later years.

Exhibit 32.1
Assessment Instruments for Older Adults

Cognitive

Screening/multiple domain: Consortium to Establish a Registry for Alzheimer's Disease; Dementia Rating Scale—2; Executive Interview; Mini-Cog; Mini-Mental State Exam; Montreal Cognitive Assessment; Neuropsychological Assessment Battery, select subtests (Judgment Test, Mazes Test, Naming Test); Repeatable Battery for the Assessment of Neuropsychological Status (RBANS); Saint Louis University Mental Status; Short Test of Mental Status; and The Neurobehavioral Cognitive Status Exam; Kokmen Short Test of Mental Status, MacNeill–Lichtenberg Decision Tree

Attention/processing speed: Digit Span, Visual Search & Attention Test

Visuospatial: Clock Drawing Test, Judgment of Line Orientation; Rey Complex Figure Test; and Trail Making Tests A and B

Language: Boston Diagnostic Aphasia Exam; Boston Naming Test; Category Fluency Test (Animals); Controlled Oral Word Association Test; Expressive Vocabulary Test—Second Edition, and Peabody Picture Vocabulary Test–Fourth Edition

Executive functioning: Brixton Test, Stroop Color–Word Test (Stroop), Wisconsin Card Sort Test (WCST); and Modified WCST

Memory: Apartment Test, Brief Visuospatial Memory Test—Revised; California Verbal Learning Test—II (CVLT–II; long or short forms); Fuld Object Memory Evaluation; Hopkins Verbal Learning Test—Revised; Memory Test for Older Adults; and Rey Auditory Verbal Learning Test

Intelligence/select domains: Test of Nonverbal Intelligence—Third Edition; Wechsler Adult Intelligence Scale—Third Edition or Fourth Edition, select subtests (Block Design, Coding, Digit Span, Similarities, Verbal tasks), Wechsler Memory Scale—Third Edition or Fourth Edition, select subtests (Logical Memory I and II, Visual Reproduction I and II), Wechsler Test of Adult Reading; and Able Reading Comprehension Test

Functional: Alzheimer's Disease Functional Assessment & Change Scale; Behavioral Dyscontrol Scale; Direct Assessment of Functional Scale (DAFS); Hopemont Capacity Assessment Interview; Independent Living Scale (ILS); Severe Impairment Battery; Functional Dementia Scale, Functional Rating Scale for the Symptoms of Dementia

Activities of daily living: Katz Index of Daily Living, Barthel Index, Instrumental Activities of Daily Living, five-item Instrumental Activities of Daily Living, Screening Questionnaire, Medical Outcomes Study Short Form, DAFS, Structured Acts of Independent Living Skill, Texas Functional Living Scale, ILS, Vineland Adaptive Behavior Scales

Capacity: The MacArthur Competence Assessment Tool; Capacity to Consent to Treatment Instrument; Hopemont Capacity Assessment Inteview, Hopkins Competency Assessment Test Capacity Evalution, Financial Capacity Instrument; Measure of Awareness of Financial Skils, Current Fianncial Capacity Form; Cognitve Capacity Screen

Effort: Medical Symptom Validity Test (MSVT), Rey-15 Item Recognition Test; Test of Memory Malingering (TOMM); Word Memory Test; Nonverbal MSVT; RBANS Effort Index, CVLT-II Forced Choice, Reliable Digit Span, Victoria Symptom Validity Test; Validity Indicator Profile; TOMM; Rey 15-Item Memory Test, Structured Inventory of Malingered Symptomatology

Mood

Adjustment: Elder Life Adjustment Interview Schedule; Affect Balance Scale, Brief Symptom Index, Subjective Happiness Scale, Subjective Well-Being Scale, Symptom Checklist—90—Revised

Depression: Cornell Scale for Depression in Dementia Scale; Geriatric Depression Scale; Zung Self-Rating Depression Scale, General Health Questionnaire, Beck Depression Inventory—II (BDI–II), Center for Epidemiologic Studies, Depression scale; Hamilton Depression Rating Scale, Patient Health Questionnaire 9—Symptom Checklist; Montgomery–Asberg Depression Rating Scale; BDI Fast Screen for Medical Patients

Anxiety: Beck Anxiety Inventory; Geriatric Anxiety Inventory; the Anxiety scale from the Hospital Anxiety and Depression Scale; State–Trait Anxiety Inventory, Trait form (STAI-T), Adult Manifest Anxiety Scale—Elderly Version; Anxiety Disorders Interview Schedule—Revised; Hamilton Anxiety Rating Scale; Penn State Worry Questionnaire; Worry Scale

Grief/bereavement: Texas Revised Inventory of Grief, Inventory of Complicated Grief

Suicide: Scale for Suicidal Ideation, Geriatric Hopelessness Scale, Beck Hopelessness Scale, Beck Scale for Suicide Ideation

Personality

NEO Personality Inventory—Revised (NEO PI–R), NEO Five-Factor Inventory (NEO FFI), and Personality Assessment Inventory (PAI), Minnesota Multiphasic Personality Inventory—2—Restructured (MMPI–2–RF)

Note. Original based on Gordon and Sweis (2009). Additional instruments from numerous sources have been added.

However, many tests continue to be used with older adults even when there is almost no research to establish the basic aspects of reliability in normative samples (e.g., Moye et al., 2006). The term *new generation instruments* is being used to explain the recently developed measures attending to extending the upper limits of normative data, but these instruments must continue to be developed to explore additional client variables such as ethnicity, medical comorbidity, and level of cognitive functioning.

CURRENT TRENDS

The U.S. Census Bureau (2008) projected that one out of five Americans will have celebrated more than 65 years of life by the year 2030, with the number of those over 65 doubling by 2050, increasing from 38.7 million to 88.5 million. Also, with the graying of America, the diversity of the population also increases, with ethnic minorities expected to become the majority by 2042 (U.S. Census Bureau, 2008).

As the growing older adult population changes the makeup of the population, so does it shape the evolving needs of psychological assessment. Because effective psychological assessment has the potential to improve the quality of care for older adults, the need to discover improved assessment practices will require limits to be pushed. Reliance must continue to be placed on long-established practices but not to the extent of constraining the evolution of psychological assessment to foster the understanding of this population.

The increasing number of older adults combined with the extended life span brings about emerging specialty areas within the areas of psychological assessment, including decision making or capacity (Moye & Marson, 2007), presence of personality disorders (Balsis, Segal, & Donahue, 2009), and the ethics of assessment in geriatric research and clinical practice (Dunn & Misra, 2009). Training may also need modification, as the necessity for assessment integration within primary care may require the development of a skill set outside of the typical range of current programs (Haley, 2005), with additional focus on specific, brevity, and normative standards.

As indicators suggest a climate moving toward accountability and best practices (Bray, 2010), evidence-based assessment is in the present and, most likely, the future. In a review of the current literature, evidence-based assessment was defined by Hunsley and Mash (2007) as

> an approach to clinical evaluation that uses research and theory to guide the selection of constructs to be assessed for a specific assessment purpose, the methods and measures to be used in the assessment, and the manner in which the assessment process unfolds. (p. 30)

A year later, these authors attempted not only to define evidence-based assessment but also to develop criteria for "assessments that work" (Hunsley & Mash, 2008). Although a work in progress, focus is being given to psychological assessment with those in their later years, and empirical knowledge for which to base best practices is gradually emerging. Perhaps there are common factors of assessment that will eventually be identified, much like Wampold et al. (1997) and others who have explored commonalities of clinical gain in psychotherapy.

Regardless of the specific course that psychological assessment in geriatric settings follows, there is no doubt that this is a rapidly growing specialty area. The future holds much opportunity and promise for older adults in need of psychological assessment as well as the psychologists who provide those clinical services.

References

American Bar Association & American Psychological Association. (2008). *Assessment of older adults with diminished capacity: A handbook for psychologists.* Retrieved from http://www.apa.org/pi/aging/capacity_psychologist_handbook.pdf

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

American Psychological Association Committee on Aging. (2009). *Multicultural competency in geropsychology*. Washington, DC: American Psychological Association.

American Psychological Association Task Force to Update the Guidelines for the Evaluation of Dementia and Age-Related Cognitive Decline. (2012). Guidelines for the evaluation of dementia and age-related cognitive change. *American Psychologist, 67*, 1–9.

American Psychological Association Working Group on the Older Adult. (1998). What practitioners should know about working with older adults. *Professional Psychology: Research and Practice, 29*, 413–427. doi:10.1037/0735-7028.29.5.413

Attix, D. K., & Welsh-Bohmer, K. A. (Eds.). (2006). *Geriatric neuropsychology: Assessment and intervention.* New York, NY: Guilford Press.

Ayis, S., Paul, C., & Ebrahim, S. (2010). Psychological disorders in old age: Better identification for better treatment. *European Journal of Psychological Assessment, 26*, 39–45. doi:10.1027/1015-5759/a000006

Baden, A. L., & Wong, G. (2008). Assessment issues for working with diverse population of elderly: Multiculturally sensitive perspectives. In L. A. Suzuki & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 594–623). San Francisco, CA: Jossey-Bass.

Baird, A. D., Ford, M., & Podell, K. (2007). Ethnic differences in functional and neuropsychological test performance in older adults. *Archives of Clinical Neuropsychology, 22*, 309–318. doi:10.1016/j.acn.2007.01.005

Balsis, S., Segal, D. L., & Donahue, C. (2009). Revising the personality disorder diagnostic criteria for the *Diagnostic and Statistical Manual of Mental Disorders—Fifth edition (DSM–V):* Consider the later life context. *American Journal of Orthopsychiatry, 79*, 452–460. doi:10.1037/a0016508

Bracken, B. A., & Reintjes, C. (2010). Age, race, and gender differences in depressive symptoms: A lifespan developmental investigation. *Journal of Psychoeducational Assessment, 28*, 40–53. doi:10.1177/0734282909336081

Bray, J. H. (2010). The future of psychology practice and science. *American Psychologist, 65*, 355–369. doi:10.1037/a0020273

Brenner, E. (2003). Consumer-focused psychological assessment. *Professional Psychology: Research and Practice, 34*, 240–247. doi:10.1037/0735-7028.34.3.240

Brick, P., Lunquist, J., Sandak, A., & Taverner, B. (2009). *Older, wiser, sexually smarter*. Morristown: Planned Parenthood of Greater Northern NJ.

Bush, S. S., & Martin, T. A. (Eds.). (2005). *Studies on neuropsychology: Neurology and cognition.* Philadelphia, PA: Taylor & Francis.

Butler, R. N. (1975). *Why survive? Being old in America.* New York, NY: Harper & Row.

Cavanaugh, J. C., Cavanaugh, C. K., Qualls, S., & McGuire, L. (Eds.). (2010). *Aging in America: Physical and mental health.* Santa Barbara, CA: Praeger.

Centers for Disease Control and Prevention and the Merck Company Foundation. (2007). *The state of aging and health in America 2007*. Whitehouse Station, NJ: The Merck Company Foundation. Retrieved from http://www.cdc.gov/aging/pdf/saha_2007.pdf

Cummins, R. A. (1997). Assessing quality of life for people with disabilities. In R. Brown (Ed.), *Quality of life for people with disabilities: Models, research and practice* (2nd ed., pp. 116–150). Cheltenham, United Kingdom: Nelson Thornes.

Dennis, R. E., Boddington, S. J. A., & Funnell, N. J. (2007). Self-report measures of anxiety: Are they suitable for older adults? *Aging and Mental Health, 11*, 668–677. doi:10.1080/13607860701529916

Duberstein, P. R., & Heisel, M. (2008). Assessment and treatment of suicidal behavior in later life. In K. Laidlaw & B. Knight (Eds.), *Handbook of emotional disorders in later life: Assessment and treatment* (pp. 311–344). New York, NY: Oxford University Press.

Dunn, L. B., & Misra, S. (2009). Research ethics issues in geriatric psychiatry. *Psychiatric Clinics of North America, 32*, 395–411. doi:10.1016/j.psc.2009.03.007

Edelstein, B. A. (2000). Challenges in the assessment of decision-making capacity. *Journal of Aging Studies, 14*, 423–437. doi:10.1016/S0890-4065(00)80006-7

Edelstein, B. A., Martin, R. R., & Koven, L. P. (2003). Psychological assessment in geriatric settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (pp. 389–414). Hoboken, NJ: Wiley.

Edelstein, B. A., Northrop, L. E., & MacDonald, L. M. (2009). Assessment. In E. Rosowsky, J. M. Casciani, & M. Arnold (Eds.), *Geropsychology and long term care: A practitioner's guide* (pp. 23–48). New York, NY: Springer. doi:10.1007/978-0-387-72648-9_3

Edelstein, B. A., Woodhead, E. L., Segal, D. L., Heisel, M. J., Bower, E. H., Lowery, A. J., & Stoner, S. A. (2007). Older adult psychological assessment. *Clinical Gerontologist, 31*, 1–35. doi:10.1080/07317110802072108

Etchells, E., Darzins, P., Silberfeld, M., Singer, P. A., McKenny, J., Naglie, G., . . . Strang, D. (1999). Assessment of patient capacity to consent to treatment. *Journal of General Internal Medicine*, *14*, 27–34. doi:10.1046/j.1525-1497.1999.00277.x

Ferraro, F. R. (2001). Assessment and evaluation issues regarding Native American elderly adults. *Journal of Clinical Geropsychology, 7*, 311–318. doi:10.1023/A:1011300309536

Fingerhood, M. (2000). Substance abuse in older people. *Journal of the American Geriatrics Society, 48*, 985–995.

Fiske, A., & O'Riley, A. A. (2008). Depression in late life. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (pp. 138–157). New York, NY: Oxford University Press.

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. *Annual Review of Clinical Psychology, 5*, 363–389. doi:10.1146/annurev.clinpsy.032408.153621

Geisinger, K. F. Boodoo, G., & Noble, J. P. (2002). The psychometrics of testing individuals with disabilities. In R. B. Ekstrom & D. K. Smith (Eds.), *Assessing individuals with disabilities in educational, employment, and counseling settings* (pp. 33–42). Washington, DC: American Psychological Association. doi:10.1037/10471-002

Gordon, B. H., & Sweis, G. W. (2009). Geriatric assessment in long term care settings: A summary of responses from LISTSERV members. *Psychologists in Long-Term Care Newsletter, 23*(2), 7–10.

Grisso, T., & Appelbaum, A. (1998). *MacArthur Competency Assessment Tool for Treatment (MacCAT-T)*. Sarasota, FL: Professional Resource Press.

Groher, M. E. (1989). Modifications in assessment and treatment for the communicatively impaired elderly. In R. H. Hull & K. M. Griffin (Eds.), *Communication disorders in aging: Sage human service guides* (pp. 103–118). Thousand Oaks, CA: Sage.

Haley, W. E. (2005). Clinical geropsychology and primary care: Progress and prospects. *Clinical Psychology: Science and Practice, 12*, 336–338. doi:10.1093/clipsy.bpi040

Halvorsrud, L., & Kalfoss, M. (2007). The conceptualization and measurement of quality of life in older adults: A review of empirical studies published during 1994–2006. *European Journal of Ageing, 4*, 229–246. doi:10.1007/s10433-007-0063-3

Heisel, M. J., & Duberstein, P. R. (2005). Suicide prevention in older adults. *Clinical Psychology: Science and Practice, 12*, 242–259. doi:10.1093/clipsy.bpi030

Helmes, E. (2000). The role of social desirability in the assessment of personality constructs. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 21–40). Norwell, MA: Kluwer Academic.

Heo, M., Murphy, C. F., Fontaine, K. R., Bruce, M. L., & Alexopoulos, G. S. (2008). Population projection of U.S. adults with lifetime experience of depressive disorder by age and sex from year 2005 to 2050. *International Journal of Geriatric Psychiatry, 23*, 1266–1270. doi:10.1002/gps.2061

Hughes, C., & Hwang, B. (1996). Attempts to conceptualize and measure quality of life. In R. L. Schalock (Ed.), *Quality of life: Conceptualization and measurement* (pp. 51–62). Washington, DC: American Association on Mental Retardation.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Mash, E. J. (Eds.). (2008). *A guide to assessments that work*. New York, NY: Oxford University Press.

Kemper, P., & Murtaugh, C. (1991). Lifetime use of nursing care. *New England Journal of Medicine, 324*, 595–600. doi:10.1056/NEJM199102283240905

Kim, H. F. S., Braun, U., & Kunik, M. E. (2001). Anxiety and depression in medically ill older adults. *Journal of Clinical Geropsychology, 7*, 117–130. doi:10.1023/A:1009585605902

Knight, B. G. (2004). *Psychotherapy with older adults* (3rd ed.). Thousand Oaks, CA: Sage.

Knight, B. G., Karel, M. J., Hinrichsen, G. A., Qualls, S. H., & Duffy, M. (2009). Pikes Peak model for training in professional psychology. *American Psychologist, 64*, 205–214. doi:10.1037/a0015059

Kogan, J. N., Edelstein, B. A., & McKee, D. R. (2000). Assessment of anxiety in older adults. *Journal of Anxiety Disorders, 14*, 109–132. doi:10.1016/S0887-6185(99)00044-4

Kresevic, D. M. (2008). Assessement of function. In E. Capezuti, D. Zwicker, M. Mezey, T. T. Fulmer, D. Gray-Miceli, & M. Kluger (Eds.), *Evidence-based geriatric nursing protocols for best practice* (3rd ed., pp. 23–40). New York, NY: Springer.

Lai, J. M., & Karlawish, J. (2007). Assessing the capacity to make everyday decisions: A guide for clinicians and an agenda for future research. *American Journal of Geriatric Psychiatry, 15*, 101–111. doi:10.1097/01.JGP.0000239246.10056.2e

Laidlaw, K., & Knight, B. (Eds.). (2008). *Handbook of emotional disorders in later life: Assessment and treatment.* New York, NY: Oxford University Press.

Lichtenberg, P. A., & Schneider, B. C. (2010). Psychological assessment and practice in geriatric rehabilitation. In R. G. Frank, M. Rosenthal, & B. Caplan (Eds.), *Handbook of rehabilitation psychology* (2nd ed., pp. 95–106). Washington, DC: American Psychological Association.

Mahalik, J. R., Good, G. E., & Englar-Carlson, M. (2003). Masculinity scripts, presenting concerns, and help seeking: Implications for practice and training. *Professional Psychology: Research and Practice, 34*, 123–131. doi:10.1037/0735-7028.34.2.123

Manly, J. J. (2006). Cultural issues. In D. K. Attix & K. A. Welsh-Bohmer (Eds.), *Geriatric neuropsychology: Assessment and intervention* (pp. 198–222). New York, NY: Guilford Press.

Merrick, E., Horgan, C., Hodgkin, D., Garnick, D., Houghton, S., Panas, L., . . . Blow, F. (2008). Unhealthy drinking patterns in older adults: Prevalence and associated characteristics. *Journal of the American Geriatrics Society, 56*, 214–223. doi:10.1111/j.1532-5415.2007.01539.x

Moberg, P. J., & Rick, J. H. (2008). Decision-making capacity and competency in the elderly: A clinical and neuropsychological perspective. *NeuroRehabilitation, 23*, 403–413.

Molinari, V., Karel, M., Jones, S., Zeiss, A., Cooley, S. G., Wray, L., . . . Gallagher-Thompson, D. (2003). Recommendations about the knowledge and skills required of psychologists working with older adults. *Professional Psychology: Research and Practice*, *34*, 435–443. doi:10.1037/0735-7028.34.4.435

Moye, J., Gurrera, R. J., Karel, M. J., Edelstein, B., & O'Connell, C. (2006). Empirical advances in the assessment of the capacity to consent to medical treatment: Clinical implications and research needs. *Clinical Psychology Review, 26*, 1054–1077. doi:10.1016/j.cpr.2005.04.013

Moye, J., & Marson, D. C. (2007). Assessment of decision-making capacity in older adults: An emerging area of practice and research. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *62*, 3–11. doi:10.1093/geronb/62.1.P3

Nuevo, R., Mackintosh, M., Gatz, M., Montorio, I., & Wetherell, J. L. (2007). A test of the measurement invariance of a brief version of the Penn State Worry Questionnaire between American and Spanish older adults. *International Psychogeriatrics, 19*, 89–104. doi:10.1017/S1041610206003450

Patterson, T. L., & Mausbach, B. T. (2010). Differences and real-world behavioral adaptation in those with mental illness. *Annual Review of Clinical Psychology, 6*, 139–154. doi:10.1146/annurev.clinpsy.121208.131339

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment, 22*, 203–212. doi:10.1037/a0018679

Qualls, S. H., & Smyer, M. A. (Eds.). (2007). *Changes in decision-making capacity in older adults: Assessment and intervention*. Hoboken, NJ: Wiley.

Reiss, N. S., & Tishler, C. L. (2008a). Suicidality in nursing home residents: Part I. *Professional Psychology: Research and Practice, 39*, 264–270. doi:10.1037/0735-7028.39.3.264

Reiss, N. S., & Tishler, C. L. (2008b). Suicidality in nursing home residents: Part II. *Professional Psychology: Research and Practice, 39*, 271–275. doi:10.1037/0735-7028.39.3.271

Rosowsky, E., Abrams, R. C., & Zweig, R. A. (Eds.). (1999). *Personality disorders in older adults: Emerging issues in diagnosis and treatment*. Mahwah, NJ: Erlbaum.

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30*, 843–861.

Schaie, K. W. (2005). *Developmental influences on adult intelligence: The Seattle Longitudinal Study*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195156737.001.0001

Schlenoff, D. (1989). Assessment of older persons with motor disabilities. In T. Hunt & C. J. Lindley (Eds.), *Testing older adults: A reference guide for geropsychological assessments* (pp. 122–134). Austin, TX: Pro-Ed.

Scott, K., Von Korff, M., Alonso, J., Angermeyer, M., Bromet, E., Bruffaerts, R., . . . Williams, D. (2008). Age patterns in the prevalence of *DSM–IV* depressive/anxiety disorders with and without physical co-morbidity. *Psychological Medicine*, *38,* 1659–1669. doi:10.1017/S0033291708003413

Segal, D. L., June, A., Payne, M., Coolidge, F. L., & Yochim, B. (2010). Development and initial validation of a self-report assessment tool for anxiety among older adults: The Geriatric Anxiety Scale. *Journal of Anxiety Disorders, 24*, 709–714.

Sharp, L. C., & Vacha-Haase, T. (2010). Physician attitudes regarding alcohol use screening in older adult patients. *Journal of Applied Gerontology*.

Stanley, M. A., & Beck, J. G. (2000). Anxiety disorders. *Clinical Psychology Review, 20*, 731–754. doi:10.1016/S0272-7358(99)00064-1

Steffens, D. C., & Potter, G. G. (2008). Geriatric depression and cognitive impairment. *Psychological Medicine, 38*, 163–175. doi:10.1017/S00332917 0700102X

Stewart, D., & Oslin, D. W. (2001). Recognition and treatment of late-life addictions in medical settings. *Journal of Clinical Geropsychology, 7*, 145–158. doi:10.1023/A:1009589706810

Substance Abuse and Mental Health Services Administration. (2009). *National survey on drug use and health*. Rockville, MD: U.S. Department of Health and Human Services.

Turk, D. C., & Okifuji, A. Skinner, M. (2008). Chronic pain in adults. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (pp. 576–591). New York, NY: Oxford University Press.

U.S. Census Bureau. (2008, August 14). *An older and more diverse nation by midcentury*. Washington, DC: U.S. Government Public Information Office.

Vacha-Haase, T., Wester, S. R., & Christianson, H. (2010). *Psychotherapy with older men.* New York, NY: Routledge.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes. *Psychological Bulletin*, *122,* 203–215. doi:10.1037/0033-2909.122.3.203

Whitbourne, S. (2000). *Psychopathology in later adulthood.* Hoboken, NJ: Wiley.

Yakushko, O., Davidson, M. M., & Williams, E. N. (2009). Identity salience model: A paradigm for integrating multiple identities in clinical practice. *Psychotherapy: Theory, Research, Practice, Training, 46*, 180–192. doi:10.1037/a0016080

CHAPTER 33

ASSESSMENT IN MARRIAGE AND FAMILY COUNSELING

Cindy I. Carlson, Lauren S. Krumholz, and Douglas K. Snyder

Many people consider having a successful marriage to be one of the most important goals in life (Roberts & Robins, 2000). Harmonious marital and family relationships, however, are not ubiquitous. Divorce, separation, and other marital or family relationship problems were the most frequently cited causes of acute emotional distress reported by adults (Swindle, Heller, Pescosolido, & Kikuzawa, 2000). Marital and family relationship distress is associated with a wide range of negative outcomes for both adults and children. Distressed couples experience increased rates of mood disorders, anxiety disorders, and substance use disorders compared with happily married couples (Whisman, 1999). Couple distress negatively affects individuals' cardiovascular, endocrine, and immune systems which, in turn, are related to physical health problems (Kiecolt-Glaser & Newton, 2001). Marital distress, conflict, and family disruption are also associated with a wide range of child problems, including depression, withdrawal, poor social competence, health problems, poor academic performance, conduct difficulties, and markedly decreased longevity (see Gottman, 1999, for a review). These data indicate that the need for couple and family counseling is high among residents of our nation.

Few people anticipate with pleasure attending marriage or family counseling. Some worry that they will be blamed for the problem; others may not view that they have a problem, are part of the problem, or can help solve it; still others may prefer avoidance as a coping mechanism. There is wide variation in the actual and perceived causes of distress in couple and family relationships as well as considerable variability in motivation to participate in counseling. The financial cost and the effort to coordinate schedules to attend therapy can also be a deterrent. Premature therapy dropout rates, in general, are high (40%–45%), which is a serious concern as research finds that early dropout eliminates treatment gains (Krumholz, 2010). Assessment of the couple or family should enhance members' engagement and motivation as well as inform treatment planning, so that early dropout among couples and families may be less likely.

Family assessment has numerous benefits and can serve multiple functions. Inclusion of a brief assessment of marital or family problems as part of a routine visit to one's primary care physician can serve to screen for relationship problems that have a strong linkage to other emotional, behavioral, and health problems (Snyder, Heyman, & Haynes, 2009). Assessment based on motivational techniques may serve to increase motivation and engagement in treatment (Dishion & Stormshak, 2006). Assessment before counseling can ensure that a broad range of routine information is collected that may be useful regarding the nature of the problem or determination of a diagnosis. Assessment should also inform treatment highlighting relationship strengths, areas of concern, possible causes of the problems, degree of distress, and variability in individual family member functioning and perspectives. A family assessment is always indicated in the psychiatric evaluation of a child or adolescent (American Academy of Child and Adolescent Psychiatry

DOI: 10.1037/14048-033
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

[AACAP], 2007) and is considered an essential step in developing a biopsychosocial treatment plan for any family member (Keitner, Heru, & Glick, 2010). Once counseling is underway, ethical practitioners continuously evaluate treatment effectiveness permitting realignment of treatment goals and strategies to serve clients better. Assessment at the end of treatment and in the follow-up phase assures that treatment goals were met and maintained.

Having established the value and many purposes served by assessment of marital and family relationships in counseling, the remainder of this chapter examines issues that are relevant to consider in conducting the assessment including guiding principles, a conceptual model, and considerations unique to marital versus family assessment. The chapter begins with a brief historical review to provide context to the discussion and ends with a review of the assessment process, ethical and legal considerations, and emerging trends.

HISTORICAL CONTEXT

Historically, marriage counseling preceded family counseling (see Nichols, 2013). The first professional centers for marriage counseling were established in the 1950s and the influential approaches of object relations marital therapy, emotionally focused couples therapy, and cognitive–behavioral marital therapy emerged. Marriage counseling focuses on the role of individual psychology in conjunction with dysfunctional relationship patterns. Because of this dual focus, marital assessment commonly includes individual health as well as dyadic relationship measures.

In contrast, assessment of family relationships has been strongly influenced by the field of family therapy (AACAP, 2007). Although the seeds of family therapy were planted in the child guidance clinics of the early 1900s, where child problems were viewed to be embedded within the family context, intervention with the family unit was uncommon until a dramatic change in thinking about the family and psychopathology emerged with Bertalanffy's *general systems theory* (Nichols, 2013). Systems theory proposed that the family, like all living systems, formed a whole comprising interrelated family members, who are reciprocally sensitive to one another and function together in a manner that attempts to maintain the homeostasis or stability of the system. From this perspective, individual problems are viewed to be a manifestation of systemic dysfunction; thus, assessment and intervention at the level of the nuclear family was preferred. Various "schools" of family therapy emerged in the 1970s and 1980s, closely followed by the development of empirically based models of family functioning and family assessment measures. Because these "first-generation" family systems models viewed psychopathology to be embedded within patterns of reciprocal family member social interactions, family assessment emphasized observation of family process. Contemporary perspectives in family counseling continue to view social interaction patterns as important in assessment; however, it is now considered important to evaluate the effect of the sociocultural community and dynamic individual developmental factors such as life experiences, internalized cognitions or narratives, and the biological/genetic bases of behavior.

GUIDING PRINCIPLES

Marital and Family Assessment Is Distinct From Assessment of Individuals

Couple and family assessment are considerably more complex than individual assessment. Assessment of the individual is widely recognized to require evaluation of multiple components, (e.g., cognitive, affective, behavioral, and physical), all of which must be considered within the larger context of influential social relationships and culture. The assessment of a marriage is at least threefold more complex because it comprises two individuals and the unique qualities and patterns of their relationship. To complicate marital assessment further, the very definition of marriage is dynamic in modern society. The contemporary dictionary definition of marriage includes both the traditional perspective, "the formal union of a man and a woman, typically recognized by law, by which they become husband and wife" and the more recent perspective, "a similar long-term relationship between partners of the same sex" (New Oxford American Online

Dictionary, 2012). Many couples that seek counseling, such as cohabiting or dating couples, however, would fit neither definition. Consistent with a broad definition, the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) includes the diagnostic category "partner relational problem" for distressed couples. Measures developed to assess a traditional conception of marriage, however, may be less valid with couples in different arrangements.

Because the numbers of individuals, possible relationships, and diversity of living arrangements is even greater in family assessment, the process is exponentially more complex. The family system is widely accepted to be an organized whole with interdependent subsystems and a hierarchical structure so that individuals are embedded within dyadic relationships that are organized by role and function (e.g., sibling, parent–child, parental, marital, grandparent), and together constitute a unique whole that is situated within an extended family system and larger sociocultural community. For purposes of illustration, in the assessment of a biological nuclear two-parent/two-child family, in addition to the assessment of four individuals, there would be assessment of the family unit as a whole, the marital relationship, the coparental relationship, the sibling relationship, four distinct parent–child relationships, and the family's interface with the larger social environment, which might include relationships with grandparents or other close extended family. Few contemporary families represent a structure as simple as this illustration!

The composition of modern American families is quite diverse and may include a host of configurations, including the traditional two-parent biological family with children; nontraditional, nonbiological parents with children; single-parent families; stepfamilies; extended kin and multigenerational families; and quasi-kin families (Bray, 2004). In addition, there is considerable variability within these categories. Families commonly extend across several households with children and adults who enter and exit on variable schedules. In recognition of contemporary family diversity, the National Institute of Mental Health (NIMH) currently defines a family as "a network of mutual commitment" (Bray, 2004, p. 11). This definition considers "family" to be those persons who currently fulfill the relationship roles traditionally specified by biological or legal relationships. Such variability of family member roles and relationships is a considerable challenge to family assessment. Who should be included in the assessment? What level of commitment meets the NIMH criteria? How valid for use with these relationships are measures and descriptive norms that have been developed primarily with two-parent biological families? In general, clinicians must be cautious about the generalizability of family measures across different social arrangements.

Assessment Must Occur on Multiple Levels

Widespread acceptance of the systems theory premise that complex systems, such as the family system, are organized hierarchically, has led to consensus among family scholars that a comprehensive assessment includes measurement at multiple system levels: (a) individual, (b) dyadic, (c) nuclear family, (d) extended family systems and related systems interfacing with the immediate family, and (e) community and cultural systems (Snyder, Cavell, Heffer, & Mangrum, 1995). Research has confirmed that different levels of family functioning, although clearly related, each contribute unique information about the family (Hayden et al., 1998).

Assessment Must Include Multiple Perspectives and Methods

Complexity of the family system, as well as the influence of systems theory in prioritizing here-and-now transactional patterns, has led family clinical scientists to recommend assessment from the perspective of both the "insider" self-reports of family members and the "outsider" observations of the couple or family members in interaction with one another. Challenges related to the valid assessment of the family as a whole using individual self-reports have also lent support to the importance of including the outsider perspective in assessment (Hampson, Beavers, & Hulgus, 1989).

Different methods are necessary to capture the insider and outsider perspectives reliably. Methods available for use in marital and family assessment

include the following: self-report questionnaires; Q-sorts; behavior ratings/checklists; projective measures; timelines of critical events; mapping and graphic techniques; structured interaction tasks; observation procedures, including formal coding of interaction and clinical rating scales; and interviews, both structured and unstructured. Many of these types of measures are described in other chapters in this handbook. Each type of family measurement method has inherent psychometric and clinical strengths and weaknesses; however, the objectivity of many methods can be enhanced with appropriate development and psychometric validation. For extended discussion of these issues, see Grotevant and Carlson (1989), and Snyder and Rice (1996).

A key distinction among family measurement methods, beyond the insider–outsider perspective, is the degree to which the method results in data that are objective; that is, numerical and systematically derived. Whereas objectivity is generally highly valued in assessment and essential to the testing of theories in research, structural adequacy in an assessment method does not assure treatment utility. Thus, the most commonly used assessment packages in marital and family assessment use some combination of the following: (a) objective self-report methods (e.g., questionnaires, behavior checklists), which are easy to administer and score; (b) clinical interviews, which may vary in subjectivity depending on the degree to which these are structured; and (c) clinical ratings of interaction, which are commonly based on observation of the couple or family engaging in structured interaction tasks.

Assessment Should Be Guided by Theory

Multiple theoretical perspectives have influenced marriage and family counseling. Each theoretical perspective has distinct assumptions and places a somewhat different emphasis on how and what to measure in an assessment.

> Ideally, any test, instrument, or procedure selected for pretreatment assessment and post-treatment outcome evaluation should operationalize one or more key concepts or constructs derived from an internally consistent and logically coherent theory of marital/family functioning, conflict/problem development, and conflict/problem resolution. (Bagarozzi & Sperry, 2004, p. 135)

Assessment Must Be Empirically Based

Clinical judgment alone is insufficient as a means of assessing couples and families. However, practitioners are urged not only to conduct assessment in marriage and family counseling but also to use empirically supported measures and methods. Measures should demonstrate psychometric quality, clinical utility, and be based on research that confirms the importance of the measured constructs to healthy couple and family functioning. One advantage of using theoretically consistent assessment methods, as have been developed within empirically based models of marital and family functioning, is the enhanced coherence and higher correlation between the insider and outsider perspectives (Hampson, Beavers, & Hulgus,1989).

Assessment Is an Ongoing Recursive Process

To guide treatment optimally, assessment and intervention processes should be continuous and interwoven throughout treatment (Snyder et al., 1995). Because time constraints in clinical practice often prohibit a comprehensive multilevel, multimethod assessment in pretreatment, Snyder et al. (1995) recommended beginning the assessment at the intermediate level of analysis; that is, at the couple level in marital counseling and at the whole family level in family counseling and then moving in both directions toward the individual level and broader system level as indicated by response to treatment. Another approach to recursive assessment and intervention is the weekly or frequent collection of assessment data. To fulfill the purpose of ongoing monitoring of treatment, a measure must be brief, easy to complete, and sensitive to change. Dishion and Stormshak (2006) used standardized parent and child daily reports weekly throughout treatment to monitor progress and needed adjustment of the intervention. Yingling (2004) reported research that documents the sensitivity to change of the Global Assessment of

Relational Functioning (GARF), which is completed by therapists, and its companion, the GARF Self-Assessment for Families (Yingling, Miller, McDonald, & Galewaler, 1998).

Ethical Practitioners Evaluate Their Clinical Work

Evaluating the effectiveness of one's clinical activities is essential to ethical clinical practice. It is important to collect data before, during, and after treatment in a follow-up phase to monitor treatment progress and to assess that a desired client outcome was achieved. Optimally, the clinician would use an A/B single-subject design, where A refers to the baseline and B refers to the treatment; data should be visually inspected for change (Jordan & Franklin, 1995).

Assessment Must Be Culturally Valid and Sensitive

Family and couples counselors rely on valid measurement and scoring procedures to plan and evaluate effective treatments. Variations in culture and history affect every domain of family functioning: composition, process, affect, and organization (Carlson, 2001). Most of the empirically based couple and family assessment models, measures, and norms, however, have been developed using European American middle-class families. The dangers of applying identical procedures and interpretative norms to cultural minority populations has been well established. Conoley and Bryant (1995) have provided examples of the cultural incongruity of numerous items selected from the most commonly used family assessment measures. Carlson (2001) concluded that attention to issues of cultural diversity in family measurement remained in its infancy. Practitioners are urged to use caution when using marital and family measures with culturally diverse families.

MARITAL AND FAMILY ASSESSMENT: A CONCEPTUAL MODEL

Snyder et al. (1995) developed a multifaceted, multilevel assessment model for evaluating couple and family distress. This model comprises five overlapping domains (i.e., cognitive, affective, behavioral, communication and interpersonal, and structural/developmental) operating at each of five system levels (i.e., individuals, dyads, the nuclear family, the extended family, and community/cultural systems). A graphic presentation of the assessment model is provided in Figure 33.1 (Snyder, Heyman, & Haynes, 2005). The conceptual model is consistent with an ecological perspective acknowledging that the marital dyad and nuclear family system are embedded within an extended family system that includes the parents' respective families of origin as well as an extended system that may include school, work, friendship, and recreational networks of social relationships, all of which are situated within a particular community and cultural niche. As illustrated in the model, information across domains may be gathered using multiple assessment strategies including both formal and informal self-report and observational techniques. Sample assessment constructs across domains and levels of couple and family functioning appear in Abbott and Snyder (2010). Snyder et al. (2009) have cautioned that this conceptual model is intended to be neither exhaustive nor prescriptive. The purpose of the model is to serve as a guiding framework for initial areas of inquiry, with the recognition that there is variation in the relevance of the particular aspects of the model for each individual couple or family. The relation between the facets of the model and relationship distress for any couple or family should be determined from an idiographic perspective (Snyder et al., 2009).

Assessing Marriages

The strength of a marriage has been conceptualized as how well a couple handles the stages of the family life cycle (AACAP, 2007). Thus, a healthy marriage involves the successful negotiation of both expected and unexpected individual and relational challenges.

Assessing the Individuals

There are strong associations between individual difficulties and couple distress. Because emotional and behavioral disorders of individuals both result from and contribute to relationship difficulties, an evaluation of individual functioning should be completed as a component of marital assessment

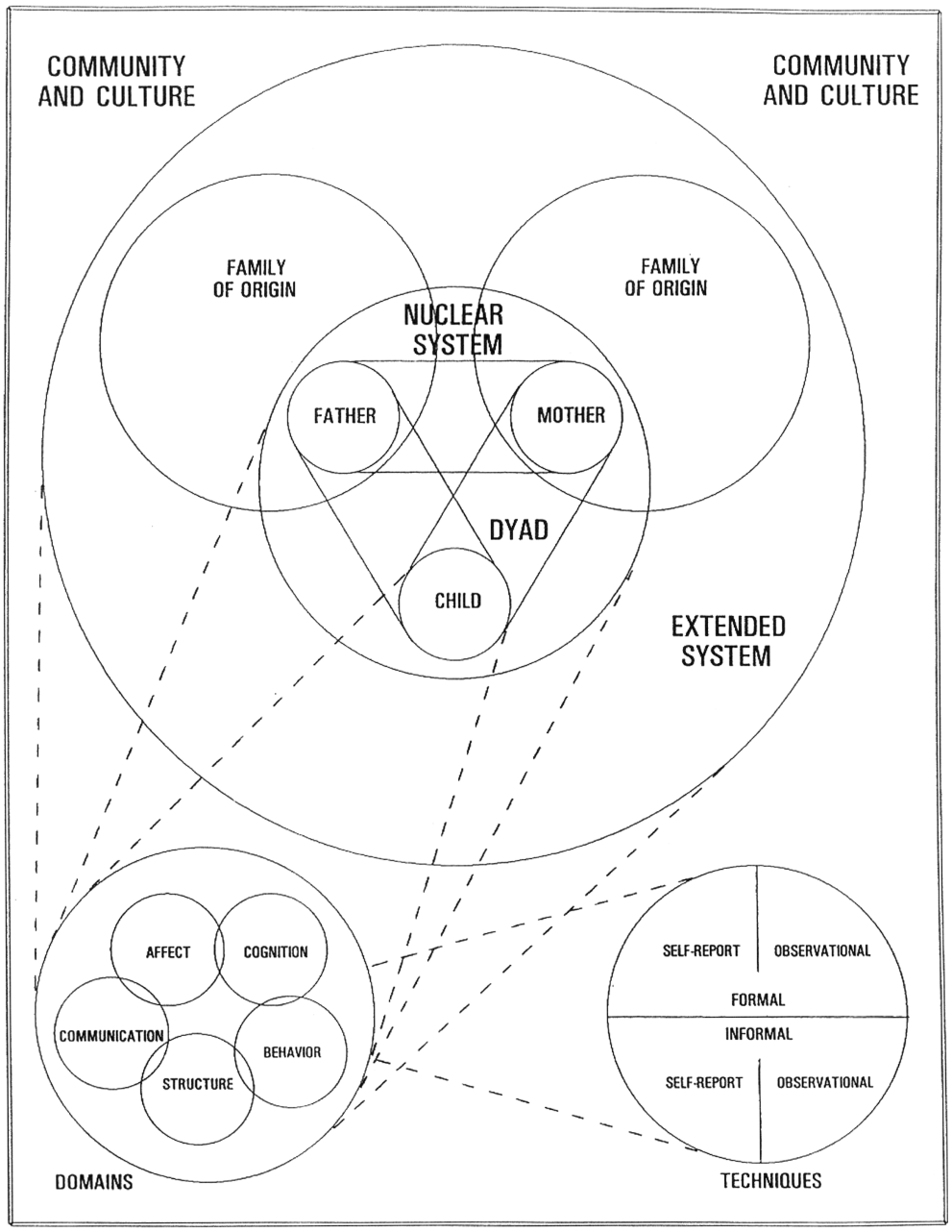


FIGURE 33.1. A conceptual model of marital and family assessment. From *Integrating Family Therapy: Handbook of Family Psychology and Systems Theory* (p. 166), by R. H. Mikesell, D. D. Lusterman, and S. H. McDaniel (Eds.), 1995, Washington, DC: American Psychological Association. Copyright 1995 by the American Psychological Association.

(Snyder & Whisman, 2003). Assessment techniques maximizing sensitivity over specificity are recommended (Snyder et al., 2005). Therapists should first screen for concerns using clinical inquiry or self-report questionnaires; when appropriate, they should follow up with measures to identify specific issues; finally, they explore how individual and relationship concerns reciprocally influence one another and are related to situational factors.

Assessing the Marital Dyad

Understanding the marital history of a couple is often the first step in assessment of the marital dyad. The marital history is an extension of the history of

each member of the couple. A carefully obtained marital history involves the collection of data on level of marital satisfaction, strengths of the marriage, and the individuals' ability to fulfill their roles. Marital history may be evaluated through the use of a variety of approaches, including interview-based methods, observational methods, and self- and other-report methods (AACAP, 2007).

Interview-based methods. The clinical assessment interview with the couple is the initial step in marital assessment and is usually conducted with both partners present. The interview has been referred to as the most versatile method of couple assessment because it can provide information across several domains and response modes (Snyder et al., 2005). For instance, it can relay useful information about particular behavioral interactions of the couple (e.g., positive and negative behavioral exchanges and problem-solving skills) as well as broader family system and cultural factors that may influence aspects of the couple's functioning. Couple interviews commonly assess the following: (a) the structure and organization of the marriage; (b) current relationship difficulties and their development; (c) previous efforts to address relationship difficulties; (d) the personality characteristics of each partner; (e) deciding whether or not to proceed with couple therapy; and (f) expectations about the therapy process (Abbott & Snyder, 2010).

Although the initial interview is central in marital assessment, and there is a strong emphasis on the use of empirically supported tools in psychological assessment, the couple assessment interview has not undergone rigorous scientific scrutiny or psychometric evaluation. Components of couple interviews, such as individual questions, have been developed and assessed; however, the psychometric properties of comprehensive structured couple assessment interviews have not been carefully and systematically studied. Nonetheless, the couple assessment interview is deemed by clinicians to be a beneficial method for identifying a couple's history and present concerns and their strengths and areas of satisfaction as well as their treatment goals, commitment to one another, and motivation for treatment.

Observational methods. The observation of a couple's interaction is a key component of marital assessment. An observational method of assessment provides direct samples of relationship behaviors in a controlled setting that are presumed to generalize to present or future behavioral patterns within the couple's natural environment (Haynes, 2001). The importance of utilizing observational techniques in the context of marital assessment should not be underestimated, as essentially all theories of relationship dysfunction and couples therapy highlight communication difficulties as a typical pathway to relational problems (Heyman, 2001). Formal and informal observational strategies have been developed to assist counselors in both hypothesis generation and hypothesis testing regarding the couples' distress in general and their communication deficits in particular. Whereas early observational coding systems were microanalytic and too time consuming for use in clinical assessment, recent adaptations provide more global and clinically useful methods. For a comprehensive review of the psychometric properties of existing observational coding systems, see Kerig and Baucom (2004).

Self- and other-report methods. The use of self-report measures has a long history within marital assessment dating back to the 1930s. The advantages of self-report measures are that they are convenient and relatively straightforward to administer while providing clinically useful information about a significant range of pertinent issues. Because of the ease of administration, they allow for the collection of large normative samples that can serve as comparison groups for clinical populations. Self-report measures also provide a means by which individuals may be able to share sensitive information that they may be uncomfortable relaying verbally. Finally, questionnaires provide insight into internal experiences and phenomena (e.g., attitudes, values, expectations, satisfaction, and commitment) that are not readily observable to an assessor. The limitations of self-report questionnaires are also noteworthy and include the following: (a) susceptibility to bias self- and other-presentation in either a favorable or unfavorable manner, (b) vulnerability to individual differences in stimulus interpretation and errors in

recollection of objective events, (c) inadvertently influencing respondents' nontest behavior in unintended ways, and (d) providing few fine-grained details concerning moment-to-moment interactions (Abbott & Snyder, 2010).

Over 1,000 self-report questionnaires that evaluate marital and family functioning have been published; however, there has not been widespread adoption of the majority of these instruments (Touliatos, Perlmutter, Straus, & Holden, 2001). Most measures have failed to demonstrate adequate reliability and validity and do not have clear evidence indicating their clinical utility (Snyder & Rice, 1996). For recommended marital assessment measures, see Abbott and Snyder (2010); Bagarozzi and Sperry (2004); and Snyder et al. (2009).

Assessing Families

Evaluation of the marital relationship is a recommended component of a family assessment; however, the focus in this section is on the additional family system levels, family roles, and extrafamilial social contexts that become relevant to assess when children are involved. Although families may seek counseling because the family, in general, experiences problems with conflict or communication, more often parents enter counseling with a desire to help a symptomatic child. Child symptoms may be biologically based but also may reflect reactions to stress in their social world. Thus, family counseling is often sought when there is a difficult family life cycle transition where child effects are a concern, such as in divorce and remarriage, military deployment of a parent, unexpected illness in the family, a geographical move, or the transition to a new level of schooling. From the perspective of the family systems practitioner, both a child referral and a family referral will point to an assessment of the whole family level of functioning, with additional levels considered as needed. When a symptomatic child is the concern, it is necessary to broaden the scope of the assessment also to include assessment of the parental subsystem, the parent–child relationship, sibling relationship, the individual child, and perspectives that reflect the child's functioning across relevant social contexts. In short, a more targeted assessment of the family unit is an excellent starting point and may be adequate when the family is the customer; when the parent is the customer seeking help for a child, a broader ecological developmental systems assessment is recommended to locate sources of stress and support. An ecological systems assessment seeks to determine in the broader social context of home, school, and neighborhood what patterns contribute to child stress and the maintenance of symptoms as well as sources of child resilience, coping, and strength.

OVERVIEW OF METHODS

Methods for assessing the family, as in marital assessment, include clinical interviews, observation methods, and self- and other-report methods. Family assessment commonly includes all three methods. Although choices may vary on the basis of the model guiding the practitioner and the age of the children, clinical interviews are most commonly conducted with the whole family to observe the interactional patterns among the family members. In addition to the clinical interview, families may be asked to engage in a series of structured tasks to elicit family interaction patterns. Structured tasks, as well as the initial clinical interview, may be evaluated using a clinical rating scale that has been developed to help systematize the observation and link data with treatment. A family assessment also typically includes completion of self- and other-report measures by family members, and an ecological assessment of the family with a symptomatic child may include other-reports and observations collected in contexts beyond the family, such as the school. The most comprehensive collection of family self-report measures appears in the *Handbook of Family Measurement Techniques* (Touliatos et al., 2001); Fischer and Cocoran (2006) have also compiled measures for clinical practice and research with couples, families, and children.

Perhaps because of the complexity and diversity of family systems as well as concerns about the cultural sensitivity of existing family assessment measures, a number of techniques have been developed that do not provide objective data regarding the family but that help the practitioner gain a "picture" of stressors, resources, and history that may inform treatment (see

TABLE 33.1

Empirical Family Assessment Models and Measures

Model	System levels assessed	Family domains assessed
Beavers Family Systems Model (Beavers & Hampson, 2000) ■ Beaver's Interactional Scales: Family Competence ■ Beaver's Interactional Scales: Family Style ■ Self-Report Family Inventory: Version II	Family	Competence Style
Circumplex Model of Marital and Family Systems/ Circumplex Assessment Package (http://www.facesiv.com) ■ Family Adaptability and Cohesion Evaluation Scales (FACES IV) ■ Clinical Rating Scales ■ Parent–Adolescent Communication Scale ■ Family Satisfaction Scale	Family Couple Parent–Adolescent Individual	Cohesion Flexibility Communication
McMaster Approach to Families (Ryan, Epstein, Keitner, Miller, & Bishop, 2005) ■ Family Assessment Device (FAD) ■ McMaster Clinical Rating Scale (MCRS) ■ McMaster Structured Interview of Family Functioning (McSiff)	Family	Problem-solving Communication Roles Affective involvement Affective responsiveness Behavior control
Process Model of Family Functioning (Skinner, Steinhauer, & Sitarenios, 2000) ■ Family Assessment Measure III (FAM-III) ■ Dyadic Relationship Scale ■ Self-Rating Scale	Family Dyadic Self	Task accomplishment Role performance Affective expression Affective involvement Control Values and norms

Note. The McMaster Family Assessment Device is available in Spanish and has been translated into many languages. The Circumplex Model of Marital and Family Systems does not have translated measurement tools in its package; however, it provides instructions for translation, and researchers have translated it into Spanish.

Thomlison, 2010, for a description). The ecomap, social network map, and social support grid are common graphic techniques that provide a depiction of the family member's contact with others, including agencies, and his or her view of these relationships as stressful or supportive. Completion of a timeline of critical events and conditions in the life cycle can place the presenting problem in a larger context that may help explain the timing of the symptom or may reveal that the symptom is part of a larger pattern of maladaptive coping. The most commonly used graphic technique, the genogram, identifies the multigenerational transmission of family patterns.

Assessing the Whole Family

Family systems theory both informs and prioritizes assessment at the whole family system level. The family systems premise of wholeness means that the family cannot be known through measurement of its individual members. As interest in family therapy grew, numerous empirically supported family models and assessment packages were developed that permit a theoretically coherent multimethod, insider–outsider assessment of the whole family system. Within each model, a self-report measure of family functioning, as well as a clinical rating scale for use in evaluating the family's interactional behavior or style, was developed. A listing of empirical family models, measures, and a comparison of their features appears in Table 33.1. Each model has amassed substantial empirical support. We next turn to a discussion and comparison of these models and methods for assessing the whole family.

Self- and other-report methods. Self-report measures are commonly used to obtain the insider perspective on family functioning. Psychometric studies of the self-report measures developed as part of the empirical models in Table 33.1 generally find high and positive correlations among them (Beavers & Hampson, 2000), suggesting considerable overlap in the domains considered relevant to assess in family processes. Similar relationships have been observed with the updated Family Adaptability and Cohesion Evaluation Scales (Olson, 2008, 2011).

One challenge to the assessment of the whole family with self-report questionnaires is the limited number and psychometric adequacy of measures appropriate for use with children younger than 10 years old. None of the empirically supported models of family functioning identified in Table 33.1 have developed child report versions, and the ability of children to make a valid assessment of marital or family relationships until recently has been questioned. Recent research suggests, however, that even young children are keen observers of family and marital interaction. Promising evidence has been provided for the use of the Berkeley Puppet Interview with children ages 4 to 7 years (see Ablow & Measelle, 2010, for a review), the GARF Self-Report for Families for children ages 8 to 12, and the SAFE cartoons for children up to 10 years old (Yingling, 2004; Yingling et al., 1998).

Observational methods. Observation of family interaction patterns is fundamental to family assessment. Valid and reliable conclusions are more easily reached by the practitioner when the observation data are systematically collected with one or more structured interaction tasks, which are videotaped and when clinical ratings are completed on videotapes of the family's interaction. Family tasks are commonly structured to elicit the following family processes: planning and problem solving; disagreement and conflict resolution; and strengths and resources.

Three of the empirical models appearing in Table 33.1 include a system by which observations of family interaction can be organized along relevant dimensions in a clinical rating scale. In a comparison of the discriminative validity of the clinical rating scales from the Beavers Family Systems Model (Beavers & Hampson, 2000), the McMaster Approach to Families (Ryan, Epstein, Keitner, Miller, & Bishop, 2005), and the Circumplex Model of Marital and Family Systems (Olson & Gorall, 2003), Drumm, Carr, and Fitzgerald (2000) found that all three rating scales correctly classified 85% of the families into clinical or nonclinical categories. Thus, all family rating scales demonstrate clinical utility, especially for screening family dysfunction. The scales varied, however, in sensitivity to particular child diagnoses (i.e., emotional, conduct, or mixed emotion–conduct disorders). The Beavers Interactional Competence Scale was best at classifying families with children with emotional disorders, and the McMaster Clinical Rating Scale (MCRS) was best at identifying families with children with mixed conduct–emotion disorders. The Circumplex Clinical Rating Scale was least sensitive in discriminating among diagnoses within the clinical group. Thus, different clinical rating scales may perform better for particular presenting child or family problems.

Interview-based methods. There are as many initial interview formats developed for use with the family in counseling as there are schools of family therapy and practitioners who publish their unique approach. By and large, practitioners are guided to use an interview method that is consistent with their theory of practice. Most models described in Table 33.1 include a structured interview that is theoretically consistent with the model and self-report family measure for enhanced coherence across perspectives.

Assessing Dyadic Relationships

The parent–child relationship. Research finds that diverse aspects of children's development and problem behavior are related to how their parents react and interact with them (Dishion & Stormshak, 2006). Thus, numerous measures of parenting and the parent-child relationship have been developed for use in research and clinical practice (see Fischer & Cocoran, 2006; Touliatos et al., 2001). Although the majority of measures are from the perspective of the parent, there has been a significant increase in measures from the perspective of the child. A commonly used measure for assessing the parent–child relationship is the Parenting Stress Index (PSI) developed by Abidin (1998). The PSI is designed

to identify potentially dysfunctional parent–child relationships among parents with children aged 3 months to 10 years. An upward extension of the measure, the Stress Index for Parenting Adolescents, is available for parents of children aged 11 to 19 years, and the Parenting Alliance Measure assesses the quality of the coparenting relationship. All instruments have strong empirical support and have been translated into numerous languages.

The sibling relationship. Although research indicates that siblings—in particular, older siblings—can play an important role in the development of problem behaviors (Dishion & Stormshak, 2006), measures to assess the sibling relationship are limited. Fine (2001) identified eight measures of sibling relations; four are appropriate for completion by elementary school-aged children, and two measures provide an assessment of the sibling relationship by parents. The clinical utility of these sibling relationship measures is largely unknown.

Assessing the Individual Child

Numerous measures have been developed to evaluate child functioning, and many are designed to capture the perspectives of parents and teachers as well as the child (see Chapter 15, this volume, and Volume 3, Chapter 6, this handbook). The family practitioner should be alert to differences in perspectives of child functioning among parents, parents and teachers, and parents and the child. Differences in these perspectives may reflect variations in behavior across contexts that provide important insight into problem solution or perceptual biases in the larger systemic context to which the child is reacting.

Assessment Across Family System Levels

One limitation to most models of family functioning is the lack of attention to self-report measurement beyond the whole family system level. For this reason, clinicians commonly must use measures developed within different models and theoretical approaches to capture the subjective perceptions of family members regarding dyadic or individual functioning. Exceptions are noteworthy. The Process Model of Family Functioning (see Table 33.1) has three versions of its self-report measure, the Family Assessment Measure III (FAM-III; Skinner, Steinhauer, & Sitarenios, 2000); each version measures identical core domains, but from the perspectives of (a) the family as a whole, (b) the dyadic family relationships, and (c) the self in the system. The advantage of this approach is the theoretical coherence across system levels. For example, the practitioner can compare different family members' perceptions of affective responsiveness in the family climate as a whole; the degree to which each member of a family dyad perceives the other member to be affectively responsive to him or her; and, finally, the individual's evaluation of his or her own affective responsiveness within the family. Although the possible data are unique with this approach, limitations include the lack of a version of the FAM-III for children below the age of 10, the completion of multiple measures may be tedious for the family, and the evidence base for this approach has not grown significantly beyond the initial development stage.

In contrast to the Process Model that assesses the same dimensions across family levels, the Darlington Family Assessment System (DFAS; Wilkinson, 1998) is one of the few that provides a differentiated evaluation of domains relevant to each system level. The DFAS assesses four key family system level perspectives deemed most important to child-centered problems: the child perspective, the parental perspective, the parent–child perspective (parenting style), and the whole family perspective. For each perspective, a set of dimensions was identified that would provide clear and meaningful distinctions among different types of problems on each systemic level. These dimensions within system levels are reflected in the Darlington Family Rating Scale, designed for use with the Darlington Family Interview Schedule. The DFAS seems to be a particularly useful organizer of the family assessment interview and postinterview data into meaningful family assessment dimensions for the practitioner; however, it has received limited empirical validation.

Family-Centered Ecological Assessment of the Child

Few family functioning models have developed measures to evaluate factors beyond the family that may contribute to child problems. On the basis of

ecological developmental and social learning theory, but also informed by family systems theory, the EcoFIT model (Dishion & Stormshak, 2006) provides an example of a structured assessment procedure that is both family centered and ecological in scope. The EcoFIT assessment includes macroratings from significant others (including parents and teachers), direct observations of family dynamics and management practices, and structured reports designed to be sensitive to change over time. The Family Assessment Task provides a series of structured interactions that are videotaped and from which a Coercive Process Index and Positive Process Index may be derived. Pretreatment baseline and change throughout treatment is measured with the Parent Daily Report and Child Daily Report, which is appropriate for completion by adolescents. Data on systems beyond the family are gathered using the Rating of Peers and Social Skills, which includes parent, child, and teacher versions. School records are commonly reviewed, and a live observation of the symptomatic child at school is conducted. The EcoFit assessment package is completed in a three-session Family Check-Up, with the third session devoted to feedback to the family. The EcoFIT approach is based on decades of research on the etiology and treatment of externalizing disorders in children and adolescents. The standard battery of instruments used in EcoFIT has been carefully researched.

THE MARITAL AND FAMILY ASSESSMENT PROCESS

In this section, we describe the marriage and family assessment process. Because the steps in psychological assessment can be similar regardless of the focus, this section provides aspects of the process unique to or emphasized in marital and family assessment. The steps in marital and family assessment may be broadly characterized as follows: (a) intake, (b) pretreatment assessment, (c) integration of data, (d) assessment feedback and collaborative treatment planning, and (e) evaluation of treatment effectiveness.

Intake

Family-centered practitioners assume that intervention begins with the initial telephone call for treatment, at which time both the intake collection of information and the preference for an initial meeting with the marital couple or family can be indicated. Clinical settings may have established intake procedures, however, that are focused on an identified patient and may pose a greater challenge to meeting with the couple or family as a unit. At a minimum, however, it can be expected that the intake worker or therapist will have had a brief telephone conversation with at least one member of the couple or family in which information can be gathered about the complaint, the household, and the motivation to change. Key questions regarding the complaint include the following: What is the nature of the concern? How long has this been a problem? What efforts have been previously made to cope with the concern? Who is involved? Beyond the family, are other persons or systems involved? Questions regarding the household include the following: Who resides in the home? What are their names, ages, genders, and relationships to one another? Are there family members who reside outside the household? Are there nonfamily members, such as the child's caretaker, who have experience with the problem? Motivation, as well as resistance to change, may quickly be revealed in discussion of who should attend therapy. Couple and family therapists, as noted, tend to prefer to see the family members together in an initial interview to permit observation of family interactions.

Pretreatment Assessment

Interview the couple/family. In the initial interview with the couple or family, the counselor will seek to (a) engage and motivate the family, (b) obtain each member's perspective on the presenting concern, (c) broaden the context of the problem through exploration of environmental and developmental stressors as well as individual and family strengths and resources, (d) observe social interaction patterns among members that may serve to perpetuate or contribute to the concern, and (e) plan for subsequent sessions. Interviews may be structured, semistructured, or open-ended (Thomlison, 2010). Structured interviews have predetermined questions; semistructured interviews use a set of questions as a beginning point from which the therapist can explore; and open-ended interviews have

no predetermined set of questions. Structured and semistructured interviews are commonly used as part of a larger marital or family assessment model or package of instruments to assure theoretical continuity across methods and measures.

Conduct standardized assessments. The purpose of using standardized assessment tools is to collect a more objective perspective on the presenting concern and family functioning. The therapist has numerous assessment methods and measures from which to choose depending on the purpose of the assessment and the presenting situation. It is appropriate to have standardized self-report measures completed by all family members for whom the reading and comprehension levels are appropriate. In conducting a family assessment where a child is the identified problem, gathering standardized measures from informed persons beyond the family, such as teachers, is recommended. Also recommended is the use of a psychometrically validated clinical rating scale to analyze the social interactions of the couple or family members. The use of structured interaction tasks or a structured interview may further enhance the reliability of data.

Integrate the Data

Whether the practitioner has completed a comprehensive assessment of the marriage and family consistent with the conceptual map provided in Figure 33.1 or a focused assessment of the subsystem levels hypothesized to be of most relevance, he or she will be faced with the challenge of integrating data from multiple perspectives. Examining data across family members, system levels, methods, and domains for convergence and divergence yields a useful depiction of the system's overall strengths and vulnerabilities. For example, in family counseling, the clinical interview may point to parental concern about a child problem; comparison of self-report data across family members may find elevated conflict scores indicating a climate of distress; and dyadic relationship data may show that the distress resides primarily in the couple relationship. Such hypotheses may be confirmed through a clinical rating of the family's behavior across several structured interaction tasks in which couple disagreement is quickly suppressed and diverted to criticism of the symptomatic child. A comparison of parent and teacher child behavior ratings may find that the parents view the child as noncompliant, whereas the teacher views the child to be more anxious and depressed. When taken together, the conclusion reached by a comparison of these assessment data is considerably different than the one that would be reached had the therapist completed only an initial interview and child behavior rating with the mother.

Synthesizing the assessment data in a family report compels the thoughtful integration of data across multiple sources and levels. Psychological assessment reports have long been used as a means by which an outside expert conducts an evaluation and communicates recommendations to others. In contrast, the contemporary perspective emphasizes the use of written information to educate, motivate, increase consumer satisfaction, gain compliance, and enhance the effectiveness of services through collaboration with the client (Wilkinson, 1998). The purpose of the report is to provide information that supports couple or family change and that places the couple or the adults in the family in the central role of change agent and decision maker (Dishion & Stormshak, 2006). Writing for this purpose necessitates that the assessment data be presented in a manner that is clearly understood by the family, the report be written without professional jargon, recommendations be based on best practices, and suggested interventions be realistic and a good match for the family within its community. Recommended family report sections include the following: overview of the family context (cultural, historical, current functioning), presenting problem, results of measures (by domain of functioning), summary of strengths and areas of concern, and recommendations (Dishion & Stormshak, 2006). These authors suggested that the recommendations section be completed after the feedback session with the couple/family, as this section of the report should reflect the collaborative discussion between the counselor and the family.

Assessment Feedback and Collaborative Treatment Planning

In summary, the integration of multilevel, multimethod data in marriage and family assessment

should clarify (a) the family's view of the problem, (b) the family's strengths and resources, and (c) the practitioner's view of the problem so that the practitioner is in a position to provide feedback to the family and collaboratively formulate an appropriate therapeutic contract. A useful model for the feedback session is provided by Dishion and Stormshak (2006), who identified four phases in the assessment feedback meeting: (a) family self-assessment, (b) therapist support and clarification, (c) therapist assessment feedback, and (d) menu of change options. Emphasized throughout the feedback session is a set of five motivational behavior change principles: data-based feedback, client responsibility for change, sound expert advice, a menu of intervention options, empathy, and the goal of enhanced self-efficacy.

Central to contemporary research on motivation is the value of choice and collaboration. Guidance for developing collaborative therapy contracts is provided by Dishion and Stormshak (2006). A collaborative therapy contract should have five components: (a) long-term goals, (b) short-term goals or immediate issues, (c) ways that improvement might first be noticed, (d) indications that goals have been achieved, and (e) plan for therapy. Making available to the family several treatment options, and discussing together the pros and cons for each option, is recommended. Once the treatment plan is agreed on verbally, the counselor should provide a written version for each party.

Evaluate Effectiveness of Treatment

Treatment effectiveness should be measured throughout treatment, at posttreatment, and at follow-up to treatment. Measures for use weekly or frequently during treatment need to be brief, related to the treatment goals, and monitor client satisfaction with the treatment. As noted earlier, excellent evidence-based options are provided by Dishion and Stormshak (2006) and Yingling (2004). Changes in the dynamics of the larger couple or family systems are best measured with a posttreatment repetition of the pretreatment assessment. This is important to ensure not only improvement in scores that had fallen within the clinical range at pretreatment but also that treatment did not result in worsened functioning within domains that were initially within normal range. Finally, although seldom completed in clinical settings, annual follow-up over time may be the most critical assessment evaluation step to perform. Krumholz (2010) found that treatment groups diverged dramatically from one another in outcomes but not until the 2nd year posttreatment.

Ethical and Legal Issues

Legal and ethical issues of confidentiality, informed consent, duty to report, and competent professional practice are related to marriage and family assessment (Thomlison, 2010). As legal requirements can vary by jurisdiction, practitioners need to learn what legal regulations apply to their assessment practice. Knowledge of family law is especially critical, given the use of assessment data in situations of family violence, child maltreatment, family dissolution, and child custody disputes. The federal law known as the Health Insurance Portability and Accountability Act, or HIPAA (PL 104-191), however, applies to all practitioners in the United States. Briefly, HIPAA permits access by the client to all his or her records and requires that written permission be obtained from the client for release of any client information, whether communicated through oral, electronic, or written means. For couple and family therapists, this requirement generally means that all legal adults in a marital or family assessment have access to the records and that written permission must be obtained from all legal adults or guardians for release of any information from an assessment performed in conjunction with marital or family counseling.

Whereas confidentiality is relatively straightforward in individual adult assessment, once again, this can be more complicated in marital or family assessment. First, there are legal limits to confidentiality, such as duty to report child abuse, child neglect, or family violence that may be more likely to emerge in a marital or family assessment. Second, family instability can result in disagreement between parents or guardians about the release of information to a third party. Before the completion of a marital or family assessment, it is important that the practitioner clearly describe the limits on confidentiality and guidelines for the release of records, such as written consent of both parents.

Informed consent, beyond issues of confidentiality, is also important in marital and family

assessment. Family members have the right to be informed about the accuracy and limitations of any assessment methods that are used as well as possible risks inherent in the assessment process. Practitioners are ethically responsible to select and rely on assessment procedures with acceptable psychometric properties, such as internal consistency, reliability, and validity.

A final ethical concern is practice within one's scope of professional competence. Marital and family assessment should be completed by appropriately trained couple and family counselors, and the completion of a couple or family assessment for use in litigation requires specialized professional knowledge, competence, and assessment procedures. The interested reader is referred to Benjamin and Gollan (2003) as well as to relevant chapters in Sperry (2004).

EMERGENT APPROACHES AND TECHNOLOGIES

Strengths-Based Assessment

Marital and family assessment has focused largely on relationship dysfunction and determining clinical levels of problem behavior in relevant domains. Problem-saturated assessments can reduce self-efficacy for change in both the counselor and the couple or family. Strengths-based assessment is a response to the limitations of problem-focused assessment. Strengths-based assessment has been defined as

> the measurement of those emotional and behavioral skills, competencies, and characteristics that create a sense of personal accomplishment; contribute to satisfying relationships with family members, peers, and adults; enhance one's ability to deal with adversity and stress; and promote one's personal, social, and academic development. (Epstein & Sharma, 1998, p. 3, cited in Rudolph & Epstein, 2000)

The development of evidence-based models of strengths-based couple and family assessment procedures is limited at present.

Therapeutic Assessment (TA)

TA is a collaborative approach to assessment in which the psychological assessment functions as a potent, short-term therapeutic intervention (for more information, see Chapter 26, this volume). Thus, TA is a semistructured hybrid of assessment and intervention strategies. Guided by consumers' questions, TA seeks to provide a therapeutic experience by collaboratively involving individuals in the assessment process and assisting them in making meaning of the assessment results by linking the findings to their everyday lives. With alterations to the model, TA has been used with families (Tharinger et al., 2009).

Assessment Technologies

Snyder et al. (2009) proposed that electronic diaries (EDs) may prove to be a useful way to collect ongoing data from couples or family members in their daily lives. With EDs, a handheld computer prompts the respondent at random or preset times to enter data about the event or treatment goal that is being monitored, such as spousal arguments or positive exchanges, along with relevant contextual information such as the time of day, social setting, antecedent exchange, and environmental stressors. The assessor later downloads stored data for real-time tabular or graphic presentation and to examine functional relations among events. Snyder et al. noted numerous benefits in the use of EDs to establish a treatment baseline and monitor goal attainment in treatment.

CONCLUDING REMARKS

Marital and family relationship distress negatively affects individual child and adult functioning. Screening for marital and family problems should be a priority in the presentation of individual child or adult symptoms, and referral for a more comprehensive assessment should be made when clinical elevations are detected. It is important to keep in mind that marital and family assessment is distinct from individual assessment and should be conducted by practitioners competent in the theory and methods relevant to marital and family counseling. In general, marriage and family counselors

prioritize assessment and intervention at the relationship level. The use of both self-report and observation methods are recommended. Marriage and family counselors have many assessment models and methods from which to choose. Ethical practitioners will prioritize use of evidence-based models and measures and will consider evaluation of one's clinical activities essential to ethical practice. Although numerous evidence-based models and measures have been developed, the marital and family counselor will continue to be challenged to identify assessment procedures that have been validated for use with children and for use with the diversity of family structures and ethnicities that populate our nation. Additionally, research has provided little guidance as to which evidence-based models and methods are preferred. Contemporary marital and family assessment encourages a more collaborative, transparent, and therapeutic approach with the consumer than was customary in previous decades. It is anticipated that this orientation, as well as greater use of technology in the collection of marital and family assessment data, will continue into the future, promising more effective and efficient monitoring of response to intervention.

References

Abbott, B. V., & Snyder, D. K. (2010). Couple distress. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (2nd ed., pp. 439–476). New York, NY: Guilford Press.

Abidin, R. (1998). *Parenting stress index (PSI)* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Ablow, J. C., & Measelle, J. R. (2010). Capturing young children's perceptions of marital conflict. In M. S. Schulz, M. K. Pruett, P. K. Kerig, & R. Parke (Eds.), *Strengthening couple relationships for optimal child development: Lessons from research and intervention* (pp. 41–57). Washington, DC: American Psychological Association. doi:10.1037/12058-004

American Academy of Child and Adolescent Psychiatry. (2007). Practice parameter for the assessment of the family. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*, 922–937.

Bagarozzi, D. A., & Sperry, L. (2004). Couples assessment: Strategies and inventories. In L. Sperry (Ed.), *Assessment of couples and families* (pp. 135–158). NY: Brunner-Routledge. doi:10.4324/9780203308271_chapter_7

Beavers, W. R., & Hampson, R. B. (2000). The Beavers systems model of family functioning. *Journal of Family Therapy, 22*, 128–143. doi:10.1111/1467-6427.00143

Benjamin, G. A. H., & Gollan, J. K. (2003). *Family evaluation in custody litigation*. Washington, DC: American Psychological Association.

Bray, J. H. (2004). Models and issues in couple and family assessment. In L. Sperry (Ed.), *Assessment of couples and families: Contemporary and cutting-edge strategies* (pp. 13–29). New York, NY: Taylor & Francis. doi:10.4324/9780203308271_chapter_2

Carlson, C. I. (2001). Family measurement overview. In J. Touliatos, B. F. Perlmutter, & G. W. Holden (Eds.), *Handbook of family measurement* (Vol. 2, pp. 1–9). Thousand Oaks, CA: Sage.

Conoley, J. C., & Bryant, L. E. (1995). Multicultural family assessment. In J. C. Conoley & E. B. Werth (Eds.), *Family assessment* (pp. 103–129). Lincoln, NE: Buros Institute of Mental Measurement.

Dishion, T. J., & Stormshak, E. A. (2006). *Intervening in children's lives: An ecological, family-centered approach to mental health care*. Washington, DC: American Psychological Association.

Drumm, M., Carr, A., & Fitzgerald, M. (2000). The Beavers, McMaster, and Circumplex clinical rating scales: A study of their sensitivity, specificity, and discriminative validity. *Journal of Family Therapy, 22*, 225–238. doi:10.1111/1467-6427.00148

Fine, M. A. (2001). Measuring family relations. In J. Touliatos, B. F. Perlmutter, & G. W. Holden (Eds.), *Handbook of family measurement* (Vol. 2, pp. 19–31). Thousand Oaks, CA: Sage Publications.

Fischer, J., & Cocoran, K. (2006). *Measures for clinical practice and research: A sourcebook.: Vol. 1. Couples, families, and children*. New York, NY: Oxford University Press.

Gottman, J. M. (1999). *The marriage clinic: A scientifically-based marital therapy*. New York, NY: Norton.

Grotevant, H. D., & Carlson, C. I. (1989). *Family assessment: A guide to methods and measures*. New York, NY: Guilford Press.

Hampson, R. B., Beavers, W. B., & Hulgus, Y. F. (1989). Insiders and outsiders' views of the family. *Journal of Family Psychology, 3*, 118–136. doi:10.1037/h0080536

Hayden, L. C., Schiller, M., Dickstein, S., Seifer, R., Miller, I., Keitner, G., . . . Rasmussen, S. (1998). Levels of family assessment: I. Family, marital,

and parent-child interaction. *Journal of Family Psychology, 12*, 7–22. doi:10.1037/0893-3200.12.1.7

Haynes, S. N. (2001). Clinical application of analogue behavioral observation: Dimensions of psychometric evaluation. *Psychological Assessment, 13*, 73–85. doi:10.1037/1040-3590.13.1.73

Health Insurance Portability and Accountability Act of 1996, 42 U.S.C. § 1320d-9 (2010).

Heyman, R. E. (2001). Observation of couple conflicts: Clinical assessment, applications, stubborn truths, and shaky foundations. *Psychological Assessment, 13*, 5–35. doi:10.1037/1040-3590.13.1.5

Jordan, C., & Franklin, C. (1995). *Clinical assessment for social workers.* Chicago, IL: Lyceum Books.

Keitner, G. I., Heru, A. M., & Glick, I. D. (2010). *Clinical manual of couples and family therapy*. Arlington, VA: American Psychiatric Publishing.

Kerig, P. K., & Baucom, D. H. (Eds.). (2004). *Couple observational coding systems.* Mahwah, NJ: Erlbaum.

Kiecolt-Glaser, J. K., & Newton, T. L. (2001). Marriage and health: His and hers. *Psychological Bulletin*, *127*, 472–503. doi:10.1037/0033-2909.127.4.472

Krumholz, L. (2010). *Maintenance of treatment effects from cognitive-behavioral therapy and parent training on family functioning and girls' depressive symptoms.* Unpublished doctoral dissertation. University of Texas, Austin.

New Oxford American Online Dictionary. (2012). *Marriage* [Online version 2.1.3]. New York, NY: Oxford University Press.

Nichols, M. P. (2013). *Family therapy: Concepts and methods* (10th ed.). Boston, MA: Pearson.

Olson, D. (2008). *FACES IV manual.* Minneapolis, MN: Life Innovations.

Olson, D. (2011). FACES IV and the circumplex model: Validation study. *Journal of Marital and Family Therapy*, *37*, 64–80. doi:10.1111/j.1752-0606.2009.00175.x

Olson, D. H., & Gorall, D. M. (2003). Circumplex model of marital and family systems. In F. Walsh (Ed.), *Normal family processes: Growing diversity and complexity* (3rd ed., pp. 459–486). New York, NY: Guilford Press.

Roberts, B. W., & Robins, R. W. (2000). Broad dispositions, broad aspirations: The intersection of personality traits and major life goals. *Personality and Social Psychology Bulletin*, *26*, 1284–1296. doi:10.1177/0146167200262009

Rudolph, S. M., & Epstein, M. H. (2000). Empowering children and families through strength-based assessment. *Reclaiming Children and Youth, 8,* 207–209.

Ryan, C. E., Epstein, N. B., Keitner, G. I., Miller, I. W., & Bishop, D. S. (2005). *Evaluating and treating families: The McMaster approach.* New York, NY: Routledge.

Skinner, H., Steinhauer, P., & Sitarenios, G. (2000). Family Assessment Measure (FAM) and process model of family functioning. *Journal of Family Therapy, 22*, 190–210. doi:10.1111/1467-6427.00146

Snyder, D. K., Cavell, T. A., Heffer, R. W., & Mangrum, L. F. (1995). Marital and family assessment: A multifaceted, multi-level approach. In R. H. Mikesell, D. D. Lusterman, & S. H. McDaniel (Eds.), *Integrating family therapy: Handbook of family psychology and systems theory* (pp. 163–182). Washington, DC: American Psychological Association. doi:10.1037/10172-009

Snyder, D. K., Heyman, R. E., & Haynes, S. N. (2005). Evidence-based approaches to assessing couple distress. *Psychological Assessment, 17*, 288–307. doi:10.1037/1040-3590.17.3.288

Snyder, D. K., Heyman, R. E., & Haynes, S. N. (2009). Assessing couples. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 457–484). New York, NY: Oxford University Press.

Snyder, D. K., & Rice, J. L. (1996). Methodological issues and strategies in scale development. In D. H. Sprenkle & S. M. Moon (Eds.), *Research methods in family therapy* (pp. 216–237). New York, NY: Guilford Press.

Snyder, D. K., & Whisman, M. A. (Eds.). (2003). *Treating difficult couples: Helping clients with coexisting mental and relationship disorders.* New York, NY: Guilford Press.

Sperry, L. (Ed.). (2004). *Assessment of couples and families.* New York, NY: Brunner-Routledge. doi:10.4324/9780203308271

Swindle, R., Heller, K., Pescosolido, B., & Kikuzawa, S. (2000). Responses to nervous breakdowns in America over a 40-year period: Mental health policy implications. *American Psychologist, 55*, 740–749. doi:10.1037/0003-066X.55.7.740

Tharinger, D. J., Finn, S. E., Gentry, L., Hamilton, A., Fowler, J., Matson, M., . . . Walkowiak, J. (2009). Therapeutic assessment with children: A pilot study of treatment acceptability and outcome. *Journal of Personality Assessment*, *91,* 238–244. doi:10.1080/00223890902794275

Thomlison, B. (2010). *Family assessment handbook* (3rd ed.). Belmont, CA: Brooks/Cole.

Touliatos, J., Perlmutter, B. F., Straus, M. A., & Holden, G. W. (Eds.). (2001). *Handbook of family measurement techniques* (Vols. 1–3). Thousand Oaks, CA: Sage.

Whisman, M. A. (1999). Marital dissatisfaction and psychiatric disorders: Results from the National

Comorbidity Survey. *Journal of Abnormal Psychology, 108*, 701–706. doi:10.1037/0021-843X.108.4.701

Wilkinson, I. (1998). *Child and family assessment: Clinical guidelines for practitioners* (2nd ed.). New York, NY: Routledge.

Yingling, L. C. (2004). Child and family assessment: Strategies and inventories. In L. Sperry (Ed.), *Assessment of couples and families* (pp. 159–181). New York, NY: Brunner-Routledge. doi:10.4324/9780203308271_chapter_8

Yingling, L. C., Miller, W. E., Jr., McDonald, A. L., & Galewaler, S. T. (1998). *GARF assessment sourcebook: Using the DSM-IV Global Assessment of Relational Functioning.* New York, NY: Brunner-Routledge.

CHAPTER 34

ASSESSMENT IN CUSTODY HEARINGS: CHILD CUSTODY EVALUATIONS

H. Elizabeth King

A child custody evaluation (CCE) is a forensic evaluation conducted to provide information and, at times, recommendations about parenting plans, physical custodial time, visiting arrangements, and decision making about the child postdivorce or after the termination of a nonmarital relationship in which a child or children were conceived. Almost 90% of parents resolve those issues without litigation, and only a small number of litigated cases involve a CCE. Disputes continue about the limited scientific basis on which a psychologist might rely to reach conclusions about such issues and, consequently, the appropriateness of the psychologist serving in such a role. Nevertheless, CCEs are nationally accepted and codified by statute or court rules in many states.

CONTEXT FOR CCES

A CCE is conducted in the context of the legal system. There is a significant difference between a clinical evaluation for which most psychologists are trained and that of a forensic evaluation. Because the legal system is adversarial, not cooperative or collaborative, the CCE must be conducted in a neutral, unbiased manner addressing any possible conflicts of interest that the evaluator might be viewed as having. Unlike other witnesses, the CCE evaluator is an expert and has the ability and competence to draw inferences from the facts that a jury, or other fact finder, cannot. There are standards that must be met in a forensic evaluation with regard to the admissibility of evidence. Most states have adopted either the *Frye* standard (*Frye v. United States*) which addresses community standards, or the *Daubert* standard (*Daubert v. Merrell Dow Pharmaceuticals*), which addresses the reliability and relevance of the evidence, peer review of the test, the test's potential error rate, its usefulness, publication of findings, and general acceptance of the technique or methodology.

GUIDELINES FOR CCES

The American Psychological Association (APA) has ethical guidelines (APA, 2010a) as well as specialized guidelines for CCEs in family law proceedings (APA, 2010b) as does the Association of Family and Conciliation Courts (AFCC; 2006). The AFCC has recommended the following:

> Prior to commencing evaluations, evaluators shall take reasonable steps to secure court orders or consent agreements (informed consent) in which they are specifically named and in which their roles, the purpose of their evaluations, and the focus of their evaluations are clearly defined. (p. 6)

The evaluator is not working for either parent or the child. It is the court to whom the evaluator is responsible and to whom the evaluator will provide his or her information about the parenting plan best suited to address the child's best interest from a psychological perspective.

DOI: 10.1037/14048-034
APA Handbook of Testing and Assessment in Psychology: Vol. 2. Testing and Assessment in Clinical and Counseling Psychology,
K. F. Geisinger (Editor-in-Chief)

BEST INTEREST STANDARD

The CCE is a complicated endeavor and requires that the evaluator be cognizant of state custody laws. Although the "best interest of the child" standard is the focus of the evaluation, the factors to be addressed vary from state to state or may not be codified. "The Michigan Standard" has been viewed as the model criteria to determine best interests of the child (Otto, Buffington-Vollum, & Edens, 2003). It includes: emotional ties between the parties and child; capacity and disposition of parties to give the child love, affection, and guidance; education and rearing of the child in a religion or creed that exists; capacity and disposition of parents to provide the child with food, clothing, medical care, or other; length of time child has lived in a stable, satisfactory environment and desirability of continuity; permanence of the proposed custodial home; moral fitness and mental and physical health of the parents; home, school, and community record of the child; reasonable preference of the child (if of sufficient age to express a preference); willingness and ability of each parent to facilitate and encourage a close and continuing parent–child relationship with the other parent; and any other factors considered by the court to be relevant to a particular child custody dispute.

PARENTING FACTORS AFFECTING CHILDREN'S ADJUSTMENT POSTDIVORCE

The effect of divorce on children has been studied extensively, and many of the factors that negatively affect children's adjustment postdivorce are well known (Amato, 2001; Amato & Keith, 1991; Emery, 2004; Emery & Forehand, 1994; Kelly, 2000; Kelly & Emery, 2003; King, 1992, 2001). Ongoing conflict between parents is the most important variable predicting emotional difficulties in children postdivorce (Amato & Keith, 1991). The importance of the role of both fathers and mothers in a child's healthy development is widely recognized; however, less clear is the necessary or sufficient amount of time spent with each parent at different levels of development to optimize healthy psychological functioning (Kelly, 2007; Kelly & Lamb, 2005; King, 2001). The negative effect on a child of a parent's having psychological problems, including depression and anxiety and/or having a personality disorder, has been documented. Furthermore, an authoritative parenting style is the most beneficial for children. *The Handbook of Parenting* (2nd ed.; Bornstein, 2002), is an excellent resource.

CCEs, by definition, involve parents in conflict who are fighting over the time (and the decision-making abilities) that they will have in the future with their child. Parents are at their worst during a divorce; therefore, many of the positive characteristics they possess will be overshadowed by the hurt, anger, and fear they are experiencing. The CCE evaluator must be aware of these issues and use past and present information to understand the family, its dynamics, and each parent's ability to understand and appropriately respond to the child's temperament, developmental needs, and problems. The child's relationship to the parents and others and the social support networks available are important considerations when assessing if a parent is capable of assisting the child's adjustment to the divorce and the resulting multiple life changes (Hetherington, 1989; Hetherington, Bridges, & Insabella, 1998; Hetherington & Kelly, 2002). There are a multitude of factors and competing needs and interests that must be assessed in the CCE. Furthermore, although the research involving a child's adjustment to different types of custody arrangements does not indicate the best arrangement for a particular child, it is important to be aware of the literature before suggesting an arrangement in a particular CCE.

Otto, Buffington-Vollum, and Edens (2003) have stressed the need for an assessment of each parent's knowledge, understanding, beliefs, values, attitudes, and behaviors pertaining to parenting and to assess the characteristics most likely to have a positive or negative effect on parenting. These traits might include the ability to empathize, follow rules, and control emotions and behaviors. Stress management skills, social skills, communication skills, and the ability to provide a good role model for a child are important to assess. Caldwell (2005) described the Big Five basic issues: quality of the attachment and bonding, potential for antisocial behavior, temper control, alienation of affection, and chemical abuse

and dependence. In some cases, more specific issues must be addressed, such as special education placement, child sexual abuse, domestic violence, and relocation that require more specialized knowledge and expertise. Understanding the child's unique characteristics including temperament, developmental needs, psychological needs, or problems that might exist is essential to assess the abilities of each parent to meet the needs of the child. Finally, the goodness-of-fit criteria are important to determine. Parents vary in their ability to parent adequately at different stages of development. Temperaments may clash, or the needs of the child may require certain parenting abilities that a particular parent may be better able to meet.

COMPONENTS OF CCE

The CCE most often includes interviews with parents and children (who are of sufficient age), observations of the child with each parent (in a free-play or directed-play context), psychological testing of the parents, interviews with third parties, and reviews of relevant information (Ackerman, 2001; Austin, 2002; Austin & Kirkpatrick, 2004; Heilbrun, 1992; Symons, 2010). Galatzer-Levy and Kraus (1999), Gould (1998, 1999), and Martindale and Gould (2004) have attempted to describe the scientific basis for CCEs systematically. CCEs are considered by some to be the most complex and time-consuming forensic evaluations (Flens, 2005; Otto, Edens, & Barcus, 2000). Each parent will present negative information about the other and will present favorable information about himself or herself. The issue of defensiveness is a complicating factor in all parts of the CCE. The parents' personal and marital histories, their concerns about each other's parenting and/or mental stability, each parent's reason for requesting a particular custody arrangement, and their plans for the future will be explored while recognizing that defensiveness is the hallmark of a CCE. The evaluator formulates multiple hypotheses about the parents, the child, and their goodness of fit and continually tests these hypotheses. The information obtained from psychological testing can be used to assess the issues raised in the case about possible pathology of either parent, to understand the parents' personality traits and characteristics, or to create new hypotheses about a parent.

USE OF TESTS IN CCE

When choosing to administer, score, and interpret a test for a CCE, APA's most current *Standards for Educational and Psychological Testing* (American Educational Research Association, APA, & National Council on Measurement in Education, 1999) should be met.[1] The psychologist must use only tests in their area of competence, avoid the choice or interpretation of tests that introduces bias, and use appropriate tests for the test taker. Norms and the relevant evidence of validity and reliability for the test and inventories used must be known. Interpretations are made within the context of test-taking behavior (fatigue or motivation), using multiple sources of convergent test and collateral data, norms, and research data. Limitations of the tests in making interpretations should be noted, and alternative explanations or hypotheses should be noted when appropriate. If using computer-generated test interpretations, the psychologist must be sufficiently knowledgeable that he or she can interpret and evaluate their quality as well as cite the source. Given the nature of the CCE and the legal standards for forensic evaluations, it is critical that information about the tests' validity, reliability, standard errors of measurement, scaling, norms, and comparability of scores be readily available for cross-examination. (Providing such information in the CCE report would make the report cumbersome and unreadable for most attorneys and judges.) Use of the standard norms as well as comparison samples (based on people involved in custody evaluations) is very important. Remarkably, in their surveys of psychologists conducting CCEs, Bow, Flens, Gould, and Greenhut (2006) found some CCE evaluators failed to use these standards; therefore, their conclusions were based on unsupportable data not admissible in court.

The use of psychological testing in CCE is frequent and long-standing; Keilin and Bloom (1986)

[1]These standards are likely to be updated regularly; however, it is unlikely that the core areas of importance will differ in the future.

surveyed 302 mental health professionals and found that testing was used about 75% of the time. The Minnesota Multiphasic Personality Inventory (MMPI) was used most often, and the Rorschach test was second. Some of the measures used (e.g., Thematic Apperception Test [TAT], Sentence Completion, House-Tree-Person, Draw a Person) lacked: standardized administration and scoring procedures, norms, reliability data and/or validity data (Anastasi, 1958; Goldstein & Hersen, 1990; Melton, Petrila, Poythress, & Slobogin, 1997). Ackerman and Ackerman's (1997) follow-up survey of psychologists conducting CCEs found that 98% reported testing adults. The MMPI or the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher et al., 2001) remained the most frequently used test, and the Rorschach remained the second most frequently used test. The Millon Clinical Multiaxial Inventory—III (MCMI–III; Millon, Davis, Millon, & Grossman, 2009), which was not reported in the Keilin and Bloom study, was used by almost one third of the psychologists. Hagen and Castagna's (2001) reanalysis of the Ackerman data reported that only the MMPI–2 usage was sufficient to meet a standard of practice.

The importance of the information obtained from testing in a CCE is a matter of debate. Brodzinsky (1993) warned against the use of tests that "have unknown predictive validity regarding issues of custody and visitation" (p. 216); he also warned against the overuse of testing results, explaining they cannot be used independently but only within a multisource, multimethod assessment. Melton et al. (1997) recommended using psychological tests when specific problems or issues are salient in the CCE. Bow and Quinnell's (2001) survey results of CCEs were similar to Lafortune and Carpenter's (1998) earlier survey. Psychological testing of the parent and child were ranked fourth and sixth (out of 10) in importance in a CCE. Although parent–child questionnaires or behavior rating scales were often used, testing of children was infrequent. Although testing is frequently a part of a CCE, its importance varies depending on the situation and the evaluator.

Otto et al. (2000) suggested a model for the selection of tests to be used in a CCE. Their list included tests being commercially available, test manual being available, having demonstrated adequate levels of reliability and validity, the test being peer reviewed, and the test being valid for the purposes for which it is used. Also, the evaluator must have the qualifications needed to administer the test. A measure of response style is important as well (Heilbrun, 1992). Criticisms of the test and its usage in the CCE context should be reviewed to explain and defend the use or lack of use of any particular test. Furthermore, use of a lengthy test "battery," which includes tests irrelevant to the custody or parenting determination, is inappropriate. Relevance and helpfulness are two essential criteria regarding test choice for a CCE.

Obviously, the various tests utilized must be administered in a standardized manner (i.e., completed in the office without input from others and in a side-by-side manner for the Rorschach), scored correctly (when possible, computer scoring through scanning test protocols is used to eliminate errors), and interpreted using the best norms and research available (Butcher & Pope, 1993). Nevertheless, the results do not provide information about the specific person being assessed; they provide information about the characteristics of a group of people obtaining the same or similar scores. The test results are best considered "hypotheses" about the parent's current functioning and should be used in that manner.

EVALUATION OF PARENTS

The MMPI–2 and the Rorschach test are the most frequently used psychological tests; however, the MCMI–III, the Parenting Stress Index (PSI), Child Behavior Check Lists (CBCLs; e.g., Achenbach and Rescorla System of Empirically Based Assessment [ASEBA; Achenbach & Rescorla, 2001]), and other instruments are frequently used. None of the manuals for these instruments address the use of the tests for custody cases. Nonetheless, all of these tests are used to (a) determine if a parent has a significant psychopathology that may have a negative effect on parenting or (b) determine the parents' personality characteristics and address ways in which these traits or characteristics might have an effect on parenting.

MMPI and MMPI–2

The MMPI–2 (Butcher et al., 2001) has long been the mainstay of evaluating parents in the CCE context (see also Chapter 11 in this volume for additional information on the MMPI–2 and other objective measures). It is the most widely used and widely researched psychological test. It is objective and can be computer scored. The first version of the MMPI (Hathaway & McKinley, 1940) included eight Clinical Scales, named for diagnostic criterion groups: Hypochondriasis (Hs; scale 1), Depression (D; scale 2), Hysteria (Hy; scale 3), Psychopathic Deviant (Pd; scale 4), Paranoia (Pa; scale 6), Psychasthenia (Pt; scale 7), Schizophrenia (Sc; scale 8), and Hypomania (Ma; scale 9). Two additional scales were later added: Social Introversion (Si; scale 10) and Masculinity-Femininity (Mf; scale 5), resulting in 10 clinical scales. The MMPI also contained a set of Validity Scales to detect problematic responding or response bias: The Cannot Say score, the Lie (L) scale, the Infrequency (F) scale, and the Correction (K) scale. The multitude of research studies that followed established empirical correlates for the scales (Butcher & Williams, 2000; Greene, 2000).

In 1982, the MMPI was revised to update the item content and create a new normative sample while ensuring that the research base of the MMPI would continue to be useful. The MMPI–2 (567 items) profile includes the original 10 clinical scales and an additional 15 content scales that have been shown to have sufficient internal consistency and test–retest reliability. Evidence of validity for the MMPI Scale scores has been gathered through empirical investigations (Pope, Butcher, & Seelen, 2006). The Koss and Butcher critical items and the Lachar–Wrobel Critical items also were retained in the MMPI–2, and 15 supplementary scales were added: Post Traumatic Stress Disorder, Marital Distress, McAndrew Alcoholism Scale—Revised, Addiction Acknowledgement Scale, Addiction Potential Scales, and Masculine Gender Role and Feminine Gender Role (Butcher et al., 2001).

Custody Litigant Comparison Sample

Although frequently used in custody evaluations (Keilin & Bloom, 1986), MMPI–2 normative data for custody litigants were not published until 1997 (Bathurst, Gottfried, & Gottfried, 1997; Butcher, 1997). Bathurst et al. noted defensive underreporting and self-favorability. Only three clinical scales had average *T* scores above the standardization mean of 50: Scale 3 (Hy; $M = 52.3$, $SD = 7.9$); Scale 4 (Pd; $M = 50.87$, $SD = 7.35$); and Scale 6 (Pa; $M = 52.4$, $SD = 9.0$). Fifty-five percent of the sample had an elevation on Scale 3 or 6, and 16% had an elevation on Scale 4. Only 2.5% ($n = 13$) had elevations into the clinical range on any clinical scale; therefore, the litigants produced clinical profiles that were healthier than those of the normative sample. The only clinical scale to be elevated with the relative frequency found in the normative sample was 6 (7.7% in the custody sample and 8% in the normative sample). As with applicants in personnel situations, parents involved in custody litigation are likely to present themselves in the most favorable light (Bagby, Nicholson, Buis, Radovanovic, & Fidler, 1999). Clearly, the context of the situation is critical to the interpretation and understanding of test scores. Bathurst et al.'s and Butcher's sample—the largest published custody sample—should be used in addition to the standardization norms when using the MMPI–2 in a CCE.

Defensiveness consistent with the custody litigant comparison sample may be context driven; therefore, it may not indicate the extreme level of defensiveness, evasiveness, and self-favorability that are suggested when compared with the nonlitigant norms. Attempts to deal with the "invalidated" or defensive MMPI–2 in a CCE range from: interpretation of scores in the lower ranges (Caldwell, 2005), using non-K-corrected clinical scales (Flens, 2006) or readministration of the test (Bagby et al., 1999). None of these attempts is supported by research; therefore if used, this lack of standardization must be reported, and the findings must be presented with appropriate caution about their limitations.

Use of MMPI–2 findings in CCEs and their implication can be found throughout the literature. For example, Caldwell (2005) described that a high Scale 4 (Pd) correlates to impaired ability to bond; therefore, such an elevation should create a "hypothesis," or concerns about bonding between that parent and the child or children. He cautioned that the context of a CCE may create a state rather

than a trait elevation; therefore, other information must be utilized to determine if there has been a history of problems with relationships or bonding and/or if this has been observed with the current parent–child relationship. (Caldwell also contended that the revision of the MMPI lowered *T* scores on Pd by 8 to 10 points; thus, even scores that are not elevated by high K scores should be interpreted.)

An elevation on a content scale is infrequent in a custody evaluation, and elevations greater than 70 on Antisocial Practices are rare in female custody litigants (1%); however, it does occur and is highly significant. Bosquet and Egeland (2000) found that mothers with elevations on Antisocial Practices of ≥ 70 were observed to be less understanding, to be more harsh, and to have more hostile parenting styles than other mothers. They were more physically and antisocially coercive and were more physically abusive. Otto et al. (2003) found that children of parents with antisocial behaviors have a variety of significant adjustment problems, including aggression and delinquency. Caldwell (2005) stated that antisocial tendencies are "like a social disease that quickly distorts family systems and damages relationships all around" (p. 90).

The MMPI—2—Restructured Form (MMPI–2–RF)

The MMPI–2–RF (Tellegen et al., 2003) used 338 items from the MMPI–2 to create a set of new scales originally developed for use as supplementary interpretive scales for the clinical scales of the MMPI–2. A set of items named Demoralization were identified and eliminated from the clinical scales (1, 2, 3, 4, 6, 7, 8, and 9) and were put into the Demoralization Scale, RCd. "Seed" scales were constructed from the eight clinical scales and correlated with the entire MMPI item pool. The restructured clinical (RC) scales were: RCd (dem) Demoralization, RC1 (som) Somatic Complaints, RC2 (lpe) Low Positive Emotions, RC3 Cynicism (cyn), RC4 (asb) Antisocial Behaviors, RC6 (per) Ideas of Persecution, RC7 (dne) Dysfunctional Negative emotions, RC8 (abx) Aberrant Experiences, and RC9 (hpm) Hypomanic Activation. Norms for the RC scales were developed using data from the MMPI–2 Restandardization Project sample. Analyses of the RC scales with existing data sets from outpatient and inpatient data pools reported equal or greater association to the behavioral correlates than the traditional clinical scales.

The MMPI–2–RF (Ben-Porath & Tellegen, 2008) was released as an updated version of the MMPI–2; however, the scales included are largely new scales without the research base of the traditional scales. A controversy among many MMPI researchers ensued, and several criticisms of the MMPI–2–RF emerged. Initial concerns involved the reduction (i.e., elimination of 229 of items [Butcher, 2011; Hathaway, 1975]). Hathaway (1975) stated,

> I, for one, would never administer only part of the test, you should be aware of how this would affect the interpretations and the consequences which you would subsequently find through your interpretation. I, for one, would never administer only part of the test. I suspect the increment of new information would fall short.

Critics contended that the MMPI–2–RF failed to improve on the original MMPI–2 (Binford & Liliquest, 2008; Gordon, 2006; Nichols, 2006; Rogers, Sewell, Harrison, & Jordan, 2006; Rouse, Greene, Butcher, Nichols, & Williams, 2008). Others contended the authors had given inadequate attention to test revision considerations (Ranson, Nichols, Rouse, & Harrington, 2009) and had used questionable psychometrics (Simms, Casillas, Clark, Watson, & Doebbeling, 2005; Rogers et al., 2006). Also of concern was the failure to provide separate gender norms in spite of evidence of gender differences on items, scales, correlates for scales (Butcher & Williams, 2009), and the literature that showed clear differences in the symptoms and behaviors of men and women assessed by personality tests (Mason, Bubany, & Butcher, 2012), with women appearing more pathological. Some researchers have found the RC scales used in the MMPI–2–RF to be less sensitive to clinical problems than the clinical scales (Butcher, 2011), and RC2 difficulty with detecting depression was attributed to construct drift by Binford and Liliquest (2008). Additionally, studies indicate that the RF scales do not measure psychopathology and antisocial behavior in criminal populations (Gancano &

Meloy, 2009; Megargee, 2006), including a recent study by McCullaugh, Pizitz, Stolberg, and Kropp (2009) of convicted stalkers who appeared quite normal on the RC scales while being detected by the Pd clinical scale.

Of equal importance, the RF scales, because of their lack of equivalence, cannot be used in conjunction with the voluminous literature developed about the MMPI clinical and content scales. Greene (2011) argued that the MMPI–2–RF "should not be conceptualized as a revised or restructured form of the MMPI–2, but as a new self-report inventory" He contended that the "'MMPI–2' in the MMPI–2–RF is a misnomer because the only relationship to the MMPI–2 is its use of a subset of the MMPI–2 item pool, its normative group, and similar validity scales" (p. 22). He viewed the MMPI–2–RF as being at a disadvantage because of the loss of the clinical scales, which means that no code type interpretations can be used as well as the loss of the content and supplementary scales that have a rich body of research. As Gass (2009) stated, "if clinicians abandon the original Clinical scales and body of code-type information, they will sacrifice the most impressive body of empirically based interpretive material ever amassed in the history of personality assessment" (p. 442).

Fake Bad Scale (FBS). The addition of the FBS to the standard scoring of the MMPI–2–RF was another source of controversy. The FBS, renamed the Symptom Validity Scale in 2008, was developed by Lees-Haley, English, and Glenn (1991) for the detection of malingering. Criticisms about the measure include item selection (rational not empirical), no established or validated guidelines for cutoffs, overlap with MMPI items related to common symptoms for physical and emotional problems, and gender differences that pathologize women. Several studies have reported gender differences in FBS scores (e.g. Butcher, Arbisi, Atlis, & McNulty, 2003; Dean et al., 2008; Greiffenstein, Fox, & Lees-Haley, 2007; Nichols, Greene, & Williams, 2009). Using the same cutoff raw scores for attribution of malingering is likely to classify more women than men in the extreme range. FBS-r, a shortened form, was included on the MMPI–2–RF. Butcher, Gass, Cumella, Kally, and Williams (2008) described the limitations of the FBS. Ben-Porath, Greve, Bianchini, and Kaufman (2009) responded to those concerns, and Williams, Butcher, Gass, Cumella, and Kally (2009) answered that response.

Although the FBS had been accepted into evidence in the past, in more recent *Frye* hearings, the FBS was excluded from the psychologist's testimony (six cases in 2007–2009) more often than admitted (four cases in 2007–2009). The controversy has been addressed in legal publications such as the *National Law Journal* and *Lawyers USA* as well as in diverse media for the general public such as the *Wall Street Journal*, *Minneapolis Star Tribune*, and *Mother Jones*.

Use of MMPI–2–RF in CCEs. The loss of the rich body of MMPI literature combined with the paucity (in comparison) of research about the RF scales, the lack of separate gender norms (raising significant questions about gender bias, particularly against mothers), and the dropping of items in Family Problems (15), Antisocial Area (20), Marital Distress (one third of the items), and Work Functioning (21) are of particular concern and make use of the MMPI–2–RF in CCEs problematic. Furthermore, the controversies surrounding the underlying assumptions in the development of the RF scales, their lack of correspondence with the original clinical scales, questions of validity, and the psychometric strengths and weaknesses of the MMPI–2 versus the RF scales will be difficult to explain and, very likely, the focus of cross-examination. On the other hand, questions about the MMPI–2 being outdated can easily be refuted. It continues to be published by Pearson and remains the most widely used clinical personality test in the world, and articles and books continue to be written about it. Given the ongoing controversy about the MMPI–2–RF and the fact that the evaluator can only use the research accumulated about the RC scales, its use in CCEs appears fraught with difficulties.

MCMI–II or MCMI–III

Neither the MCMI–II nor the MCMI–III was reported in the Keilin and Bloom (1986) survey of psychological testing in child custody evaluations; however, it was used by one third of the psychologists in the Ackerman and Ackerman (1997) survey.

The MCMI was developed for use in treatment planning. Those taking the test are assumed to have difficulties, and the test is designed to describe their pathology. It has 10 personality disorder scales corresponding to the 10 specific personality disorders in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) and four additional personality disorder scales for entities that are not found in the *DSM–IV*. The MCMI used a clinical population as its norm group; therefore, it does not compare the client with the general population. For this reason, concerns about its applicability to a normal population have been raised. Because the MCMI was designed to *describe* problems, not to determine if they exist, its use with nonclients is likely to result in overpathologizing them. Additional questions have been raised about base rate and gender bias (Flens, 2006; McCann et al., 2001) as well as response style issues (McCann et. al., 2001). Validity issues are a significant concern as well (Craig, 2006; McCann, 2002). For example, Millon et al. (1997) explained that the Histrionic (4), Narcissistic (5), and Compulsive (7) Personality Disorder scales may at times reflect personality strengths as well as pathology. Craig commented, "research evidence suggests that these scales may not be measuring the disorders implied by their scale designations" (pp. 71–72).

Use of MCMI in CCEs

Concerns about using the MCMI in CCEs are numerous. The use of a test which overpathologizes nonpatient litigants has been raised by several researchers (Lampel, 1996; Otto & Butcher, 1995; see also Chapters 11 and 24, this volume). The usefulness of the MCMI in this context was questioned by Halon (2001) who noted that the elevation of Y (social desirability) and three (4, 5, and 7) personality disorder scales was so frequent in CCEs that he named them "the Normal Quartet in child custody cases." In their 2001 article, McCann et al. had similar findings. Further, they noted that these scales correlate in the positive direction with measures of emotional health and in the negative direction with measures of psychological disturbance (Craig, 2006); therefore, they cannot be interpreted as an indication of psychopathology. McCann et al. suggested an adjustment in scores might be needed for interpreting scales 4 and 7 in mothers or other female litigants; however, Hynan (2004) cautioned there was no solid empirical basis for this type of adjustment, and "practitioners should exercise a great deal of caution about using those scales at all" (p. 109).

The revision of the MCMI–III (Millon et al., 2009) resulted in additional concern about its use in CCEs (Flens, 2010). Although providing new norms in 2008, the norm group was smaller than the original, a single combined gender norm group was used, and no data were provided comparing the 1997 norms with the 2008 norms. The addition of the Grossman Facet Scales was questioned because of the lack of studies about the scales' reliability and validity or research supporting its use. In summary, the MCMI–III has multiple problems that preclude its use in CCEs. There is every reason to avoid using this test in a CCE.

Rorschach

The Rorschach test was originally considered a projective test until John Exner (1991) developed the Comprehensive System (CS), a variety of scoring and interpretive systems. The CS was developed in an empiricist spirit, using the most reliable, valid scores and indices from older systems and creating new ones. The result was a complex scoring and interpretive system. Significant criticisms about the CS's normative sample, reliability of scoring, and validity of interpretation were raised by Garb, Wood, Lilienfeld, and Nezworski (2005; see also Chapter 10, this volume).

In contrast, research reviews by Viglione and Hilsenroth (2001) and Weiner (2001) reported that CS stability was adequate or better when compared with other personality tests. The norms underlying the CS were criticized because of the small sample, redundancy of subjects, and inaccurate reference population (Garb et al., 2005; Wood, Nezworski, Garb, & Lilienfeld, 2001); however, the norms generally have been supported by a cross-cultural normative study (Shaffer, Erdberg, & Meyer, 2007). Wood also contended that the norms overpathologized (Wood et al., 2001); however, others have countered this argument (Weiner, 2008; Weiner & Meyer, 2009). There have been a multitude of

studies supporting Rorschach validity around the world (Exner & Erdberg, 2005). Meyer and Archer (2001) compared the evidence for the Rorschach in the context of evidence from meta-analyses on other psychological tests and concluded that their construct validity was supported. The usefulness of the Rorschach has been demonstrated by Erard (2007); Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999); Meyer (2000); and Meyer and Archer (2001). Research reviews (Viglione & Hilsenroth, 2001; Weiner, 2001) concluded there were empirical data demonstrating that the Rorschach variables possess incremental validity over other tests, including self-report scales, and two meta-analyses have reported that the predictive power of the Rorschach is comparable with that of other personality assessment measures (Gronnerod, 2004; Hiller et al., 1999). In 2005, the Society for Personality Assessment issued an Official Statement about the Rorschach, which concluded,

> The Rorschach possesses documented reliability and validity similar to other generally accepted test instruments used in the assessment of personality and psychopathology and that its responsible use in personality assessment is appropriate and justified. (p. 221)

Nevertheless, the peer-reviewed evidence for many individual CS scores and indices including the Egocentricity Index and the Depression Index is weak, and many believe that the normative base needs adjustment. No changes or developments in the CS have been permitted since Exner's death, and as a result, members of Exner's Rorschach Research Council (Greg Meyer, Donald Viglione, Phil Erdberg, Joni Mihura) and Robert Erard have developed a new evidence-based system called the Rorschach Performance Assessment System (Viglione, Meyer, & Mihura, 2010). Obviously, it is important to follow the developments in the use of the Rorschach if planning to use it in a CCE.

Rorschach in court. The introduction of Rorschach findings in court has been seriously questioned; however, Meloy, Hansen, and Weiner (1997) found that the Rorschach was used in 247 court cases between 1945 and 1995, and it had been accepted into evidence without challenge 97% of the time. Meloy's (2008) subsequent survey found that, between 1996 and 2005, the Rorschach was challenged by attorneys in only 2% of 150 cases.

Use of the Rorschach in CCEs. The Rorschach can often detect serious pathology that may have been missed on the MMPI (Ganellen, 1996); therefore, it may provide incremental validity to the CCE. It can also provide information in the areas of emotion, thinking, coping styles, interpersonal information, impulse control, self-perception, and situational stress (Calloway, 2005) and can be very helpful in understanding the parents' predilection toward cooperative, competitive, or avoidant styles of interacting. Schultz (Evans & Schultz, 2008) described four relevant issues about which the Rorschach can provide useful information in CCEs: (a) the continuum between narcissistic self-involvement and empathy; (b) boundary regulation, that is, differentiation versus enmeshment; (c) parental responsiveness (support, protection, other-focusedness vs. rejection, criticalness, disdain, and contempt); and (d) accuracy of attribution toward other people versus interpersonal distortion. An excellent resource regarding use of the Rorschach is CCEs is the *Handbook of Forensic Rorschach Assessment* (Gancano & Meloy, 2007).

In spite of the controversy about the Rorschach, it appears reasonable, if the evaluator has an acceptable level of training and experience in its use, to use the Rorschach in CCEs. On a practical level, use of the Rorschach has declined and is likely to continue to do so because the cost to cover the extensive time required for its administration, scoring, and interpretation is seldom covered by health insurance companies, and many young psychologists have not received training about the Rorschach while in graduate school.

Parenting Surveys

The Parenting Stress Index (PSI). The PSI (Abidin, 1998) is a self-report measure requiring a fourth- or fifth-grade education. It can be used with parents of children between 3 months and 12 years of age. It was developed to identify

parents who needed support and guidance, possible dysfunctional parent–child relationships, and children at risk for problems. The 101 items fall into two domains: Child Characteristics and Parent Characteristics. There are six subscales of Child Characteristics. Four address child temperament and learned behavior: adaptability, distractibility or hyperactive-type behaviors, demandingness, and mood. The remaining two relate to the cognitive–affective responses of the parent to the child: acceptability and reinforcing. The Parent Characteristics domain has seven subscales about the characteristics of the parent and his or her perception of his or her social support system for parenting: competence, isolation, attachment, health, role restriction, depression, and spouse. The norm group was 2,633 mothers. (Although the manual presents data from 200 fathers of children ages 6 months to 6 years, these were not included in their analysis.) The PSI Total, Child Domain, and Parent Domain scores have an alpha reliability between .90 and .95, but scores for the specific subscales are less robust. Abidin, Flens, and Austin (2006) provided standard errors of measurement and confidence intervals that were lacking in the PSI manual (Abidin, 1995) to facilitate interpretations. In spite of the lack of father norms and the limited size of the norm group, they argued that research indicates that parental gender differences are negligible (Deater-Deckard & Scarr, 1996) or are consistent with traditional parenting roles (Krauss, 1993). Furthermore, the research with other populations and in other countries (Bigras, Lafreniere, & Dumas, 1996; Hutcheson & Black, 1996) suggests that the PSI is reliable and would be valid for use with parents in the United States.

Custody litigant data. Abidin et al. (2006) provided preliminary data on the PSI in a custody context. They found no difference between the male and female litigants on domains, subscales, and total scores. Only the Defensive Indicator scores showed a statistical difference between male and female caretakers, with males being more defensive. Roughly one third of the sample was defensive; therefore, caution must be used when interpreting the data. Furthermore, Abidin et al. (2006) found that the defensive responding in their custody group was strongly related to self-deceptive enhancement, which, they discussed, indicates that these parents have "a narcissistic overconfidence in their abilities, have a lack of insight into their own behavior, and are easily angered when confronted" (p. 318).

Use of PSI in CCEs. Given the PSI scores' association with insecure attachment, quality of parent–child interactions, parents' perceptions about the child, harshness of punishments, depressive parent behaviors, and likely types of problem behaviors of children, it is useful as a source of data about the parent–child relationships and the types of stressors in the family. Its emphasis on attributes of the parent and child that might lead to dysfunctional parenting (Abidin, 1995) makes it particularly helpful in a CCE, but the lack of normative data for fathers is problematic. As with most of the other tests used in a CCE, the defensiveness of parents involved in custody litigation requires that caution be used in interpreting the data. Use of the preliminary custody litigant data in addition to the traditional norms is important.

Use of the PSI in a custody modification in which one or both of the parents has remarried is problematic because of the Spouse Subscale, which is likely to elicit information about the current spouse rather than the other biological parent; therefore, Abidin et al. (2006) now has recommended that, in these situations, the parent should be given instructions that the spouse questions refer to the other biological parent of the child. Using the altered instructions appears most appropriate in the CCE context; however, it is important that a caveat be made in the report.

EVALUATION OF THE CHILD

This section reflects the assessment of a child or children in a traditional CCE. If there are specific concerns about a child, such as questions of retardation, a learning disability, attention problems, or emotional difficulties, a more focused evaluation of the child is required and additional tests are likely to be used.

The ASEBA (Achenbach & Rescorla, 2001) includes the following measures: CBCLs for Children ages 6 to 18 (CBCL/6-18); the Teacher's Report

Form (TRF); and the Youth Self-Report (YSF). The CBCL/6-18 is a revision of the CBCL/4-18 (Achenbach, 1991; Achenbach & Edelbrock, 1983). The YSR is normed for children ages 11 to 18 and has children describe themselves. The TRF is completed by teachers and other school personnel. The ASEBA was designed to capture ways in which children behaved in different situations, including with different parents. These instruments provide standardized ratings and other descriptive information about the child and the child's self-view. The normative sample of 1,285 included 11- to 18-year-olds who had no known mental health or substance abuse problems. Reliability (test–retest of .79–.90) and validity data (content criterion-related, and construct evidence of validity) are provided in the ASEBA manual (Achenbach & Rescorla, 2001).

Tests Administered to Children and Adolescents

Both Lafortune and Carpenter's (1998) and Bow and Quinnell's (2001) surveys reported testing of children was infrequent in CCEs. This is undoubtedly because children's scores on personality tests often lack evidence of reliability and validity (Flens, 2006). There are three notable exceptions: The Minnesota Multiphasic Personality Inventory—Adolescent (MMPI–A; Butcher et al., 1992), the YSF (Achenbach, 1991), and the Rorschach. The YSF is used with children ages 11 to 18; the Rorschach, with children ages 5 to 17; and the MMPI–A, with adolescents ages 12 to 17. The YSF and the MMPI–A are easily administered and computer scored. They can provide specific information about the child or adolescent's attitudes and/or problems.

Other Measures

Some CCE evaluators find it useful to use cards from the TAT or a Sentence Completion Test to elicit content themes from children. Drawings also may be used with children as a way to engage them with a nonthreatening task while eliciting their feelings, understanding, and concerns about their parents' divorce, their perceptions of each parent and/or the purpose of the CCE. The child also may be asked to draw pictures of the family or activities. If useful information is elicited ("Mommy hides her vodka under the bed," "I hate it when Daddy hits Mommy at night," or "I wish Mommy wouldn't hit me so much"), this is reported as part of the child interview but no testing conclusions are made.

Custody-Specific Tests

Several tests were designed specifically to assist in determining custody appropriateness. The Bricklin Perceptual Scales (BPS; Bricklin, 1990), the Ackerman–Schoendorf Scales for Parent Evaluation of Custody (ASPECT; Ackerman & Schoendorf, 1992), and A Comprehensive Custody Evaluation Stand System (ACCESS; Bricklin & Elliott, 1995) are the best known.

The BPS consists of 32 questions (16 for each parent) that are asked of the child, who provides a verbal and a nonverbal response (poking a hole in a black line anchored by "very well" or "not so well"). Shaffer (1992) criticized both the theories underlying the assumption that the child's nonverbal response reflected their unconscious attitudes as well as the meager statistics presented. Hagin (1992) criticized the small sample size (12–36) and the lack of descriptive statistics presented about the sample.

The ASPECT incorporates some commonly used instruments, including the MMPI–2 and Rorschach, clinical data, and data from questionnaires completed by both parents that yield a Parental Custody Index (PCI). Otto and colleagues (Otto et al., 2003; see also Otto et al., 2000) had concerns about the ASPECT's basic psychometric properties and the use of judges' decisions as a criterion for predictive validity. Melton (1995) criticized the ASPECT for inadequate psychometric construction, lack of data regarding its reliability and validity, and the lack of an empirical research base. Arditti's (1995) concerns included cumbersome administration, poor scale reliability, absence of validity data, and the small normative sample (100 participants). Otto et al. (2000) echoed the concern about the lack of validity data. Melton also noted the lack of any methodologically sound research published in refereed scientific journals to support the use of the ASPECT in CCEs. Ackerman (2005) and Connell (2005) engaged in a point–counterpoint about the ASPECT in the book *Psychological Testing in Child Custody Evaluations*.

The ACCESS (Bricklin & Elliott, 1995) utilizes several measures, including the BPS. Others include the Parent Awareness Skills Survey, which has no information about reliability or validity in the manual and which Bischoff (1995) concluded was necessary before this instrument was used clinically. The Parent Perception of Child Profile (Bricklin & Elliott, 1991) has no information on scoring, norms, reliability, or validity (Kelley, 1995); and the Perception of Relationships Test has no existing norms (Carlson, 1995), reliability data, and little validity support (Conger, 1995). Clearly, the ACCESS does not meet minimum standards for test construction, reliability, or validity and does not have an adequate research base.

Despite being developed for use in CCEs to assist in the determination of a parent's appropriateness for being a custodial parent, all of these systems are seriously flawed and have insufficient reliability and validity data, and they lack norms and an empirical research base. Their decline in usage is likely the result of these difficulties (Connell, 2005a, 2005b; Gould, 2005) and their failure to meet the standards required for use in a CCE.

Computerized Interpretations

Computer-generated reports are available for many of the tests: the MMPI–2, the MMPI–2–RF, the Millon, the Rorschach, and the PSI. Their use has been discussed at length (Butcher, 2003; Butcher, Perry, & Dean, 2009; Fowler, 1969; Moreland, 1985; Otto & Butcher, 1995; Otto & Collins, 1995; Otto et al., 2000). Because the algorithms on which they are based are "trade secrets," their use in forensic work has been questioned. Certainly, the user should obtain the research studies on which the algorithms are based and should be able to interpret tests without the assistance of the computerized interpretation. Appropriate use of computer-generated interpretations allows the evaluator to have a backup with regard to their interpretations and a check on their interpretations in that the computer reports are based only on research findings; therefore, they are not vulnerable to bias. Furthermore, the computer, unlike a human, does not forget; therefore, it can provide detailed descriptive information about the scores that the clinician may have forgotten.

CONCLUSION

At the present time, the MMPI–2 is the only psychological test meeting the required psychometric standards and being used with sufficient frequency to meet the standard of commonly accepted for use in CCEs, and it is consistently admitted into evidence. The Rorschach is used with less frequency and remains controversial; however, its use can provide incremental data and is most often admitted into evidence. The MMPI–2–RF's problems detecting depression and antisocial personalities combined with the controversies about gender bias and the concerns regarding the FBS make it a dubious choice for CCEs. The MCMI–III's problems with validity and gender bias rule out its use in a CCE. The PSI, with limitations that have been described, is the only parenting survey with adequate norms and an adequate research base for use in a CCE. The Achenbach measures have sufficient psychometric strength and meet the standard of usage; however, there are no custody-litigant data for children or parents. Tests specifically designed for custody determination lack sufficient reliability and validity data to be used in CCEs. Obviously, other measures such as the TAT, House-Tree-Person, or Incomplete Sentences should never be considered tests. They may be used and reported in the interview sections of a report, but caution should be taken to clarify their use as only a clinical measure.

When using test results in a CCE, it is imperative to clarify that the results do not describe the individual being assessed; instead, the results are a description of the personality characteristics of a group of people. It remains the task of the evaluator to obtain the necessary information to confirm or disconfirm the "hypotheses" indicated by the test data. To the extent that other data support the hypotheses, this should be conveyed in the report. Furthermore, it is imperative that information contrary to the test data be reported as well.

Information and clinical impressions made during the parent interviews and information obtained from other sources are combined with the results of psychological tests to provide the psychologist a picture of the parent's personality strengths and weaknesses that have already affected or potentially

might affect their parenting. As Martindale (2005) described, evaluators must be mindful of the risk of confirmatory bias, which is the finding and reporting of information supportive of one's favored hypotheses. He offered the term *confirmatory distortion* as the intentional engagement of selective reporting or skewed interpretation of data to bolster a favored hypothesis. The evaluator in a CCE is not immune to such problems and must work to avoid them. Martindale has suggested being careful not to be influenced by a very likeable litigant, make premature conclusions about the litigants, become an interested party, or become ego involved. He discusses the dangers of selective recall, the effect of primacy, and the evaluator's inability to discern deceit by verbal or nonverbal indicators (Feeley & Young, 1998). These pitfalls are some of the reasons why it is important to conceptualize multiple hypotheses about the case and look for nonconfirmatory data as well as supporting data. The evaluator's final hypothesis or assessment of the parties must be supported by the data and be sufficiently strong to withstand rigorous adversarial questioning about the contradictory data.

There is no established weight given to any type or source of information when conducting a CCE or making conclusions about the parents, child, and the parenting arrangement most likely to be in the best interests of the child, and is unlikely that such a formula will ever be successfully devised. For example, the test data and information from the child may not play a significant role in the evaluation of a litigant whose interview elicits multiple behaviors that suggest antisocial practices and who lies to the evaluator during the interview (documented by third-party information such as police reports, court transcripts, etc.). In another situation, testing may be very helpful in raising questions about a parent's knowledge of and ability to care for a child. For example, the parent who responds to the feelings rather than the behavior of a child who is conduct disordered may generate concerns about his or her ability to provide appropriate structure and consequences to the child. When ASEBA ratings from the four sources can be obtained and compared, important information about each parent's knowledge of their child can be obtained. A parent whose ratings about a child indicate no difficulties—when the child's TRF, YSF, and other parent's CBCL ratings indicate affective issues—may be unable to accept the child's emotional issues and fail to provide the emotional support and/or interventions needed. In another situation, a significant history of depression combined with current difficulties in managing the behavior of the child or children may override normal test data and information that the parent is not currently depressed. Every effort should be made by the evaluator to use consistent methods of evaluation in CCEs; however, the unique qualities of the individual parents and child demand different weight or priorities be given to the information obtained in each CCE. This complex process should be clearly spelled out in the report so that the Court might determine if the assessment has been conducted in an unbiased and useful manner. The evaluator's responsibilities include clarifying the data and the logic used in making the conclusions reached about the parents and the child and the data supporting or disputing these and the logic used in any recommendations made. Failure to do so results in additional costs to the litigants in terms of time and expense, since depositions and so forth will be needed to elicit this information. Experts disagree on how much supporting research should be included in the CCE report, but all agree that the evaluator must have this information available should the attorneys or the court question their decision making. The task of the evaluator is to provide the court with sufficient information to make it obvious that a realistic and unbiased assessment has been made of each parent's strengths and limitations in parenting.

References

Abidin, R. R. (1995). *Parenting Stress Index: Professional manual* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Abidin, R. R. (1998, August). Parenting Stress Index: Its empirical validation. In *Symposium on child custody.* Symposium conducted at the 106th Annual Convention of the American Psychological Association, San Francisco, CA.

Abidin, R. R., Flens, J. T., & Austin, W. G. (2006). The parenting stress index. In R. Archer (Ed.), *Forensic uses of clinical assessment instruments* (pp. 297–328). Mahwah, NJ: Erlbaum.

Achenbach, T. M. (1991). *Manual for the Child Behavior Check List/4–18 and 1994 Profile*. Burlington: Department of Psychiatry, University of Vermont.

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Check List 4–18 and Revised Child Behavior Profile*. Burlington: Department of Psychiatry, University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington: Research Center for Children, Youth, and Families, University of Vermont.

Ackerman, M. J. (2001). *Clinician's guide to child custody evaluations* (2nd ed.). New York, NY: Wiley.

Ackerman, M. J., (2005). The Ackerman–Schoendorf Scales for Parent Evaluation of Custody (ASPECT): A review of research and update. In J. Flens & L. Drozd (Eds.), *Psychological testing in child custody evaluations* (pp. 179–194). New York, NY: Haworth Press.

Ackerman, M. J., & Ackerman, M. C. (1997). Child custody evaluation practices: A survey of experienced professionals (revisited). *Professional Psychology: Research and Practice, 28*, 137–145. doi:10.1037/0735-7028.28.2.137

Ackerman, M. J., & Schoendorf, K. (1992). *ASPECT: Ackerman–Schoendorf Scales for Parent Evaluation of Custody—Manual*. Los Angeles, CA: Western Psychological Services.

Amato, P. R. (2001). Children and divorce in the 1990s: An update of the Amato and Keith (1991) meta-analysis. *Journal of Family Psychology, 15*, 355–370. doi:10.1037/0893-3200.15.3.355

Amato, P. R., & Keith, B. (1991). Parental divorce and the well being of children: A meta-analysis. *Psychological Bulletin*, *110*, 26–46. doi:10.1037/0033-2909.110.1.26

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

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

American Psychological Association. (2010a). *Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010)*. Retrieved from http://www.apa.org/ethics/code/index.aspx

American Psychological Association. (2010b). Guidelines for child custody evaluations in family law proceedings. *American Psychologist, 65*, 863–867. doi:10.1037/a0021250

Anastasi, A. (1958). *Psychological testing* (6th ed.). New York, NY: Macmillan.

Arditti, J. A. (1995). Ackerman–Schoendorf Scales for Parent Evaluation of Custody. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 20–22). Lincoln, NE: Buros Institute of Mental Measurements.

Association of Family and Conciliation Courts. (2006). *Model standards of practice for child custody evaluation*. Retrieved from http://www.afccnet.org/Portals/0/ModelStdsChildCustodyEvalSept2006.pdf

Austin, W. G. (2002). Guidelines for utilizing collateral sources of information in child custody evaluations. *Family Court Review, 40*, 177–184. doi:10.1111/j.174-1617.2002.tb00828.x

Austin, W. G., & Kirkpatrick, H. D. (2004). The investigation component in forensic mental health evaluations: Considerations in the case of parenting time evaluations. *Journal of Child Custody, 1*, 23–46. doi:10.1300/J190v01n02_02

Bagby, R. M., Nicholson, R. A., Buis, T., Radovanovic, H., & Fidler, B. J. (1999). Defensive responding on the MMPI-2 in family custody and access evaluations. *Psychological Assessment, 11*, 24–28. doi:10.1037/1040-3590.11.1.24

Bathurst, K., Gottfried, A., & Gottfried, A. (1997). Normative data for the MMPI-2 in child custody litigation. *Psychological Assessment, 9*, 205–211. doi:10.1037/1040-3590.9.3.205

Ben-Porath, Y. S., Greve, K. W., Bianchini, K. J., & Kaufmann, P. M. (2009). The MMPI-2 Symptom Validity Scale (FBS) is an empirically validated measure of over-reporting in personal injury litigants and claimants: Reply to Butcher et al. (2008). *Psychological Injury and Law*, *2,* 62–85. doi:10.1007/s12207-009-9037-4

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI–2–RF: Manual for administration, scoring and interpretation*. Minneapolis: University of Minnesota Press.

Bigras, M., Lafreniere, P., & Dumas, J. (1996). Discriminant validity of the parent and child scales of the Parenting Stress Index. *Early Education and Development, 7*, 167–178. doi:10.1207/s15566935eed0702_5

Binford, A., & Liliquest, L. (2008). Behavioral correlates of selected MMPI–2 clinical, content, and restructured clinical scales. *Journal of Personality Assessment, 90*, 608–614. doi:10.1080/00223890802388657

Bischoff, L. G. (1995). Review of the parent awareness skills survey. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 735–736). Lincoln, NE: Buros Institute of Mental Measurements.

Bornstein, M. (Ed.). (2002). *The handbook of parenting* (2nd ed.). Mahwah, NJ: Erlbaum.

Bosquet, M., & Egeland, B. (2000). Predicting parenting behaviors from antisocial practices: Content

scale scores of the MMPI–2 administered during pregnancy. *Journal of Personality Assessment, 74*, 146–162. doi:10.1207/S15327752JPA740110

Bow, J. N., Flens, J. R., & Gould, J. W. (2010). MMPI–2 and MCMI–III in forensic evaluations: A survey of psychologists. *Journal of Forensic Psychology Practice*, *10,* 37–52. doi:10.1080/15228930903173021

Bow, J. N., Flens, J. R., Gould, J. W., & Greenhut, D. (2006). An analysis of administration, scoring, and interpretation of the MMPI–2 and MCMI–III in child custody evaluations. *Journal of Child Custody: Research, Issues, and Practices*, *2,* 1–22.

Bow, J. N., & Quinnell, F. A. (2001). Psychologists' current practices and procedures in child custody evaluations: Five years after American Psychological Association guidelines. *Professional Psychology: Research and Practice, 32*, 261–268. doi:10.1037/0735-7028.32.3.261

Bricklin, B. (1990). *Bricklin Perceptual Scales manual.* Furlong, PA: Village.

Bricklin, B., & Elliott, G. (1991). *Parent Perception of Child Profile.* Furlong, PA: Village.

Bricklin, B., & Elliott, G. (1995). *ACCESS: A comprehensive custody evaluation standard system.* Furlong, PA: Village.

Brodzinsky, D. M. (1993). On the use and misuse of psychological testing in child custody evaluations. *Professional Psychology: Research and Practice, 24*, 213–219. doi:10.1037/0735-7028.24.2.213

Butcher, J. N. (1997). *Frequency of MMPI–2 scores in forensic evaluations*. Retrieved from http://www1.umn.edu/mmpi/documents/Frequency/MMPI2Scores.pdf

Butcher, J. N. (2003). Computer based psychological assessment. In J. Graham, & J. Naglieri (Eds.), *Comprehensive handbook of psychology: Vol 10. Assessment psychology* (pp. 141–164). New York, NY: Wiley.

Butcher, J. N. (2011). *A beginner's guide to the MMPI–2* (3rd ed.). Washington, DC: American Psychological Association.

Butcher, J. N., Arbisi, P. A., Atlis, M. M., & McNulty, J. L. (2003). The construct validity of the Lees-Haley Fake Bad Scale: Does this scale measure somatic malingering and feigned emotional distress? *Archives of Clinical Neuropsychology, 18*, 473–485.

Butcher, J. N., Gass, C. S., Cumella, E., Kally, Z., & Williams, C. L. (2008). Potential for bias in the MMPI–2 assessments using the Fake Bad Scale (FBS). *Psychological Injury and the Law, 1*, 191–209. doi:10.1007/s12207-007-9002-z

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, Y. S., Dahlstrom, W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory—2: Manual for administration and scoring* (rev. ed.). Minneapolis: University of Minnesota Press.

Butcher, J. N., Perry, J., & Dean, B. L. (2009). Computer-based assessment. In J. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 163–182). New York, NY: Oxford University Press.

Butcher, J. N., & Pope, K. S. (1993). Seven issues in conducting forensic assessments: Ethical responsibilities in light of new standards and new tests. *Ethics and Behavior*, *3,* 267–288.

Butcher, J. N., & Williams, C. L. (2000). *Essentials of the MMPI–2 and MMPI–A clinical interpretation* (2nd ed.). Minneapolis: University of Minnesota Press.

Butcher, J. N., & Williams, C. L. (2009). Personality assessment with the MMPI-2: Historical roots, international adaptations, and current challenges. *Applied Psychology: Health and Well-Being*, *1,* 105–135. doi:10.1111/j.1758-0854.2008.01007.x

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Daemmer, B. (1992). *MMPI–A manual for administration, scoring and interpretation.* Minneapolis: University of Minnesota Press.

Caldwell, A. B. (2005). How can the MMPI–2 help child custody examiners? *Journal of Child Custody*, *2,* 83–117. doi:10.1300/J190v02n01_06

Calloway, G. C. (2005). The Rorschach: Its use in child custody evaluations. *Journal of Child Custody*, *2,* 143–157. doi:10.1300/J190v02n01_08

Carlson, J. F. (1995). Perception of relationships test. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 746–747). Lincoln, NE: Buros Institute of Mental Measurements.

Conger, J. (1995). Review of Perception-of-Relationships Test. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 747–748). Lincoln, NE: Buros Institute of Mental Measurements.

Connell, M. (2005). Review of "The Ackerman–Schoendorf Scales for Parent Evaluation of Custody (ASPECT)." In J. Flens & L. Drozd (Eds.), *Psychological testing in child custody evaluations* (pp. 195–210). New York, NY: Haworth Press.

Craig, R. J. (2006). The Millon clinical multiaxial inventory-III. In R. Archer (Ed.), *Forensic uses of clinical assessment instruments* (pp. 121–145). Mahwah, NJ: Erlbaum.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579. (1993).

Dean, A. C., Boone, K. B., Kim, M. S., Curiel, A. R., Martin, D. J., Victor, T. L., . . . Lang, Y. K. (2008). Examination of the impact of ethnicity on the Minnesota Multiphasic Personality Inventory—2

(MMPI–2) Fake Bad Scale. *The Clinical Neuropsychologist, 22*, 1054–1060. doi:10.1080/13854040701750891

Deater-Deckard, K., & Scarr, S. (1996). Parenting stress among dual-earner mothers and fathers: Are there gender differences? *Journal of Family Psychology, 10*, 45–59. doi:10.1037/0893-3200.10.1.45

Emery, R. (2004). *The truth about children and divorce: Dealing with the emotions so you and your children can survive.* New York, NY: Viking/Penguin.

Emery, R. E., & Forehand, R. (1994). Parental divorce and children's well-being: A focus on resilience. In R. Haggerty, L. Sherrod, N. Garmezy, & M. Rutter (Eds.), *Stress, risk, and resiliency in children and adolescents* (pp. 64–99). Cambridge, England: Cambridge University Press.

Erard, R. E. (2007). Picking cherries with blinders on: A comment on Erickson et al. (2007) regarding the use of tests in family court. *Family Court Review*, *45*, 175–184. doi:10.1111/j.1744-1617.2007.00137.x

Evans, F. B., & Schultz, B. M. (2008). The Rorschach in child custody and parenting plan evaluation: A new conceptualization. In C. B. Gancoco & B. B. Evans (Eds.), *The handbook of forensic Rorschach assessment*. New York, NY: Routledge.

Exner, J. E. (1991). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (3rd ed.). New York, NY: Wiley.

Exner, J. E., & Erdberg, P. E. (2005). *The Rorschach: A comprehensive system: Vol. 2. Advanced interpretation*. Hoboken, NJ: Wiley.

Feeley, T. H., & Young, M. J. (1998). Humans as lie detectors: Some more second thoughts. *Communication Quarterly, 46*, 109–126.

Flens, J. R. (2005). The responsible use of psychological testing in child custody evaluations: Selection of tests. In J. Flens & L. Drozd (Eds.), *Psychological testing in child custody evaluations* (pp. 3–29). New York, NY: Haworth Press.

Flens, J. R. (2006, October). *Psychological testing in child custody evaluations.* Advanced workshop presented at the Association of Family and Conciliation Courts Conference, Atlanta, GA.

Flens, J. R. (2010, March). *Use of the MMPI–2 and MCMI–III in forensic practice: What's all the fighting about?* Paper presented at the meeting of the Society for Personality Assessment, Boston, MA.

Fowler, R. D. (1969). Automated interpretation of personality test data. In J. Butcher (Ed.), *MMPI: Research developments and clinical applications.* New York, NY: McGraw-Hill.

Frye v. United States, 293 F. 1013 D. C. Cir. 1923.

Galatzer-Levy, R. M., & Kraus, L. (Eds.). (1999). *The scientific basis of child custody decisions.* New York, NY: Wiley.

Gancano, C. B., & Meloy, R. R. (2007). *Handbook of forensic Rorschach assessment.* New York, NY: Routledge.

Gancano, C. B., & Meloy, R. R. (2009). Assessing antisocial and psychopathic personalities. In J. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (2nd ed., pp. 361–375). New York, NY: Oxford University Press.

Ganellen, R. J. (1996). *Integrating the Rorschach and MMPI–2 personality assessment.* Mahwah, NJ: Erlbaum.

Garb, H. N., Wood, J. M., Lilienfeld, S. O., & Nezworski, T. (2005). Roots of the Rorschach controversy. *Clinical Psychology Review, 25*, 97–118. doi:10.1016/j.cpr.2004.09.002

Gass, C. S. (2009). Use of the MMPI–2 in neuropsychological evaluations. In J. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 432–456). New York, NY: Oxford University Press.

Goldstein, G., & Hersen, M. (Eds.). (1990). *Handbook of psychological assessment* (pp. 387–399). New York, NY: Pergamon Press.

Gordon, R. M. (2006). False assumptions about psychopathology, hysteria and the MMPI–2 restructured clinical scales. *Psychological Reports, 98*, 870–872. doi:10.2466/pr0.98.3.870-872

Gould, J. W. (1998). *Conducting scientifically crafted child custody evaluations.* Thousand Oaks, CA: Sage.

Gould, J. W. (1999). Scientifically crafted child custody evaluations: II: A paradigm for forensic evaluation of child custody determination. *Family Court Review*, *37,* 159–178. doi:10.1111/j.174-1617.1999.tb00534.x

Gould, J. W. (2005). Use of psychological tests in child custody assessment. *Journal of Child Custody*, *2,* 49–69. doi:10.1300/J190v02n01_04

Greene, R. L. (2000). *MMPI–2/MMPI: An interpretive manual* (2nd ed.). Boston, MA: Allyn & Bacon.

Greene, R. L. (2011). *MMPI–2/MMPI–2–RF: An interpretive manual* (3rd ed.). Boston, MA: Allyn & Bacon.

Greene, R. L., Rouse, S. V., Butcher, J. N., Nichols, D. S., & Williams, C. L. (2009). The MMPI–2 Restructured Clinical (RC) scales and redundancy: Response to Tellegen, Ben-Porath, and Sellbom. *Journal of Personality Assessment*, *91*, 222–226. doi:10.1080/00223890902800825

Greiffenstein, M. F., Fox, D., & Lees-Haley, P. R. (2007). The MMPI–2 Fake Bad Scale in detection of noncredible brain injury claims. In K. Boone (Ed.), *Detection of noncredible cognitive performance* (pp. 210–235). New York, NY: Guilford Press.

Gronnerod, C. (2004). Rorschach assessment of changes following psychotherapy: A meta-analysis review.

Journal of Personality Assessment, 83, 256–276. doi:10.1207/s15327752jpa8303_09

Hagen, M. A., & Castagna, N. (2001). The real numbers: Psychological testing in child custody evaluations. *Professional Psychology: Research and Practice*, *32,* 269–271. doi:10.1037/0735-7028.32.3.269

Hagin, R. A. (1992). Review of the Bricklin Perceptual Scales. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 117–118). Lincoln, NE: Buros Institute of Mental Measurements.

Halon, R. L. (2001). The Millon Clinical Multiaxial Inventory—III: The normal quartet in child custody cases. *American Journal of Forensic Psychology*, *19,* 57–75.

Hathaway, S. R. (1975, February). Comment on MMPI abbreviated forms [Tape recording]. In J. N. Butcher (Chair), *Who owns test items? Present confusions and anxieties about 1984.* Symposium conducted on the Recent Developments in the Use of the MMPI, St. Petersburg, FL.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. *Journal of Psychology, 10*, 249–254. doi: 10.1080/00223980.1940.9917000

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257–272. doi:10.1007/BF01044769

Hetherington, E. M. (1989). Coping with family transitions: Winners, losers, and survivors. *Child Development, 60*, 1–14. doi:10.2307/1131066

Hetherington, E. M., Bridges, M., & Insabella, G. M. (1998). What matters? What does not? Five perspectives on the association between marital transitions and children's adjustment. *American Psychologist, 53*, 167–184. doi:10.1037/0003-066X.53.2.167

Hetherington, E. M., & Kelly, J. (2002). *Divorce reconsidered: For better or for worse.* New York, NY: Norton.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11*, 278–296. doi:10.1037/1040-3590.11.3.278

Hutcheson, J. J., & Black, M. M. (1996). Psychometric properties of Parenting Stress Index in a sample of low income African-American mothers of infants and toddlers. *Early Education and Development, 7*, 381–400. doi:10.1207/s15566935eed0704_5

Hynan, D. J. (2004). Unsupported gender differences on some personality disorder scales of the Millon Clinical Multiaxial Inventory—III. *Professional Psychology: Research and Practice*, *35,* 105–110. doi:10.1037/0735-7028.35.1.105

Keilin, W. G., & Bloom, L. J. (1986). Child custody evaluation practices: A survey of experienced professionals. *Professional Psychology: Research and Practice, 17*, 338–346. doi:10.1037/0735-7028.17.4.338

Kelley, M. L. (1995). Review of the parent perception of child profile. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 738–739). Lincoln, NE: Buros Institute of Mental Measurements.

Kelly, J. B. (2000). Children's adjustment in conflicted marriage and divorce: A decade of review of research. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 963–973. doi:10.1097/00004583-200008000-00007

Kelly, J. B. (2007). Children's living arrangements following separation and divorce: Insights from empirical and clinical research. *Family Process*, *46,* 35–52. doi:10.1111/j.1545-5300.2006.00190.x

Kelly, J. B., & Emery, R. (2003). Children's adjustment following divorce: Risk and resilience perspectives. *Family Relations, 52*, 352–362. doi:10.1111/j.1741-3729.2003.00352.x

Kelly, J. B., & Lamb, M. E. (2001). Using child development research to make appropriate custody and access decisions for young children. *Family Court Review*, *39,* 249–266. doi:10.1111/j.174-1617.2001.tb00609.x

King, H. E. (1992). The reactions of children to divorce. In C. Walker & M. Roberts (Eds.), *Handbook of clinical child psychology* (2nd ed., pp. 1009–1023). New York, NY: Wiley.

King, H. E. (2001). Children and divorce. In C. Walker & M. Roberts (Eds.), *Handbook of clinical child psychology* (3rd ed., pp. 1031–1045). New York, NY: Wiley.

Krauss, M. W. (1993). Child-related and parenting stress: Similarities and differences between mothers and fathers of children with disabilities. *American Journal on Mental Retardation, 97*, 393–404.

Lafortune, L. A., & Carpenter, B. N. (1998). Custody evaluations: A survey of mental health professionals. *Behavioral Sciences and the Law, 16*, 207–224. doi:10.1002/(SICI)1099-0798(199821)16:2<207::AID-BSL303>3.0.CO;2-P

Lampel, A. K. (1996). Children's alignment with parents in highly conflicted custody cases. *Family Court Review*, *34,* 229–239. doi:10.1111/j.174-1617.1996.tb00416.x

Lees-Haley, P. R., English, L. T., & Glenn, W. J. (1991). A fake bad scale on the MMPI–2 for personal injury claimants. *Psychological Reports, 68*, 203–210. doi:10.2466/pr0.1991.68.1.203

Martindale, D. A. (2005). Confirmatory bias and confirmatory distortion. *Journal of Child Custody*, *2,* 31–48. doi:10.1300/J190v02n01_03

Martindale, D. A., & Gould, J. W. (2004). The forensic model: Ethics and scientific methodology applied

to custody evaluations. *Journal of Child Custody*, *1,* 1–22. doi:10.1300/J190v01n02_01

Mason, S. N., Bubany, S., & Butcher, J. N. (2012). *Frequently asked questions: Gender differences on personality tests*. Retrieved from http://www1.umn.edu/mmpi/documents/Gender%20Differenes%20in%20personality%20FREQUENTLY%20ASKED%20QUESTIONS.pdf

McCann, J. T. (2002). Guidelines for forensic application of the MCMI–III. *Journal of Forensic Psychology Practice*, *2,* 55–69. doi:10.1300/J158v02n03_04

McCann, J. T., Flens, J. R., Campagna, V., Colman, P., Lazzaro, T., & Connor, E. (2001). The MCMI–III in child custody evaluations: A normative study. *Journal of Forensic Psychology Practice, 1*, 27–44. doi:10.1300/J158v01n02_02

McCullaugh, J. M., Pizitz, T. D., Stolberg, R., & Kropp, J. (2009, March). *A comparison study between the MMPI–2–RF profiles of convicted stalkers*. Presented at the meeting of the Society for Personality Assessment, Chicago, IL.

Megargee, E. I. (2006). *Using the MMPI–2 in criminal justice and correctional settings*. Minneapolis: University of Minnesota Press.

Meloy, J. R. (2008). The authority of the Rorschach: An update. In C. Gacono & F. Evans (Eds.), *The handbook of forensic Rorschach psychology* (pp. 79–88). New York, NY: Routledge.

Meloy, J. R., Hansen, T. L., & Weiner, I. B. (1997). Authority of the Rorschach: Legal citations during the past 50 years. *Journal of Personality Assessment, 69*, 53–62. doi:10.1207/s15327752jpa6901_3

Melton, G. (1995). Ackerman–Schoendorf Scales for Parent Evaluation of Custody. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 22–23). Lincoln, NE: Buros Institute of Mental Measurements.

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (1997). *Psychological evaluations for the courts* (3rd ed.). New York, NY: Guilford Press.

Meyer, G. J. (2000). On the science of Rorschach research. *Journal of Personality Assessment*, *75*, 46–81. doi:10.1207/S15327752JPA7501_6

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13*, 486–502. doi:10.1037/1040-3590.13.4.486

Michigan Child Custody Act. Act 91 of 1970.

Millon, T., Davis, R. M., Millon, C., & Grossman, S. (2009). *Millon Clinical Multiaxial Inventory—III* (4th ed.). Bloomington, MN: Pearson Clinical Assessment.

Millon, T., Millon, C., & Davis, R. M. (1997). *MCMI–III manual* (2nd ed.). Minneapolis, MN: National Computer Systems.

Moreland, K. L. (1985). Validation of computer based interpretations: Problems and prospects. *Journal of Consulting and Clinical Psychology, 53*, 816–825. doi:10.1037/0022-006X.53.6.816

Nichols, D., Greene, R., & Williams, C. L. (2009, March). *Gender bias in the MMPI–2 Fake Bad Scale (FBS) and FBS-r in MMPI–2–RF*. Paper presented at the meeting of the Society for Personality Assessment, Chicago, IL.

Nichols, D. S. (2006). The trials of separating bath water from the baby: A review and critique of the MMPI-2 Restructured Clinical scales. *Journal of Personality Assessment, 87*, 121–138. doi:10.1207/s15327752jpa8702_02

Otto, R. K., Buffington-Vollum, J. K., & Edens, J. F. (2003). Child custody evaluation. In J. Weiner (Series Ed.) & A. Goldstein (Vol. Ed.), *Handbook of psychology: Vol. 2. Forensic psychology* (pp. 179–208). New York, NY: Wiley.

Otto, R. K., & Butcher, J. N. (1995). Computer-assisted psychological assessment in child custody evaluations. *Family Law Quarterly*, *29,* 79–92.

Otto, R. K., & Collins, R. P. (1995). Use of the MMPI–2/MMPI–A in child custody evaluations. In S. Hobfoll (Series Ed.), & Y. Ben-Porath, J. Graham, G. Hall, R. Hirschman, & M. Zaragonza (Vol. Eds.), *Applied psychology: Individual, social, and community issues: Vol. 2. Forensic applications of the MMPI–2* (pp. 222–252). New York, NY: Wiley.

Otto, R. K., Edens, J. F., & Barcus, E. H. (2000). The use of psychological testing in child custody evaluations. *Family Court Review, 38*, 312–340. doi:10.1111/j.174-1617.2000.tb00578.x

Pope, K. S., Butcher, J. N., & Seelen, J. (2006). Assessing malingering and other aspects of credibility. In K. Pope, J. Butcher, & J. Seelen (Eds.), *The MMPI, MMPI–2, and MMPI–A in court: A practical guide for expert witnesses and attorneys* (3rd ed., pp. 129–160). Washington, DC: American Psychological Association. doi:10.1037/11437-007

Ranson, M., Nichols, D. S., Rouse, S. V., & Harrington, J. (2009). Changing or replacing an established personality assessment standard: Issues, goals, and problems, with special reference to recent developments in the MMPI–2. In J. Butcher (Ed.), *Handbook of personality assessment* (pp. 112–139). New York, NY: Oxford University Press.

Rogers, R., Sewell, K. W., Harrison, K. S., & Jordan, M. J. (2006). The MMPI–2 Restructured Clinical scales: A paradigmatic shift in scale development. *Journal of Personality Assessment, 87*, 139–147. doi:10.1207/s15327752jpa8702_03

Rouse, S. V., Greene, R. L., Butcher, J. N., Nichols, D. S., & Williams, C. L. (2008). What do the MMPI–2 Restructured Clinical scales reliably measure?

Journal of Personality Assessment, 90, 435–442. doi:10.1080/00223890802248695

Shaffer, M. B. (1992). Review of the Bricklin Perceptual Scales. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 118–119). Lincoln, NE: Buros Institute of Mental Measurements.

Shaffer, T. W., Erdberg, P., & Meyer, G. J. (Eds.). (2007). International reference sample for the Rorschach comprehensive system [Special issue]. *Journal of Personality Assessment, 89*(3).

Simms, L. J., Casillas, A., Clark, L. A., Watson, D., & Doebbeling, B. N. (2005). Psychometric evaluation of the Restructured Clinical scales of MMPI–2. *Psychological Assessment, 17*, 345–358. doi:10.1037/1040-3590.17.3.345

Society for Personality Assessment. (2005). The status of the Rorschach in clinical and forensic practice: An official statement by the Board of Trustees of the Society for Personality Assessment. *Journal of Personality Assessment, 85*, 219–237. doi:10.1207/s15327752jpa8502_16

Symons, D. K. (2010). A review of the practice and science of child custody and access assessment in the United States and Canada. *Professional Psychology: Research and Practice, 41*, 267–273. doi:10.1037/a0019271

Tellegen, A., Ben-Porath, Y. S., McNulty, J., Arbisi, P., Graham, J. R., & Kaemmer, B. (2003). *MMPI–2: Restructured Clinical (RC) Scales*. Minneapolis: University of Minnesota Press.

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13*, 452–471.

Viglione, D. J., Meyer, G. J., & Mihura, J. L. (2010, March) *Using emerging and existing data to improve Rorschach validity and utility*. Workshop presented at the Annual Meeting of the Society of Personality Assessment, San Jose, CA.

Weiner, I. B. (2001). Considerations in collecting Rorschach reference data. *Journal of Personality Assessment*, *77*, 122–127. doi:10.1207/S15327752JPA7701_08

Weiner, I. B. (2008). Presenting and defending Rorschach testimony. In C. B. Gancono & F. B. Evans (Eds.), *The handbook of forensic Rorschach assessment* (pp. 121–140). New York, NY: Routledge.

Weiner, I. B., & Meyer, G. I. (2009). Personality assessment with the Rorschach inkblot method. In J. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 277–298). New York, NY: Oxford University Press.

Williams, C., Butcher, J. N., Gass, C. S., Cumella, E., & Kally, Z. (2009). Inaccuracies about the MMPI–2 Fake Bad Scale in the reply by Ben-Porath, Greve, Bianchini, and Kaufman. *Psychological Injury and Law, 2*, 182–197. doi:10.1007/s12207-009-9046-3

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001). The misperception of psychopathology problems with the norms of the comprehensive system. *Clinical Psychology: Science and Practice, 8*, 350–373. doi:10.1093/clipsy.8.3.350